

ABSTRACT

Title of Dissertation: THREE VARIATIONS OF PRECISION
MEDICINE: GENE-AWARE GENOME
EDITING, ANCESTRY-AWARE
MOLECULAR DIAGNOSIS, AND CLONE-
AWARE TREATMENT PLANNING

Sanju Sinha, Doctor of Philosophy, 2021

Dissertation directed by: Dr. Eytan Ruppín, Department of Computer
Science; Prof. Steve Mount, Department of
Biology

During my Ph.D., I developed several computational approaches to advance precision medicine for cancer prevention and treatment. My thesis presents three such approaches addressing these emerging challenges by analyzing large-scale cancer omics data from both pre-clinical models and patients datasets.

In the first project, we studied the cancer risk associated with CRISPR-based therapies. Therapeutics based on CRISPR technologies (for which the chemistry Nobel prize was awarded in 2020) are poised to become widely applicable for treating a variety of human genetic diseases. However, preceding our work, two experimental studies have reported that genome editing by CRISPR–Cas9 can induce a DNA damage response mediated by p53 in primary cells hampering their growth. This could lead to an undesired selection of cells with pre-existing p53 mutations. Motivated by these findings, we conducted the first comprehensive computational and experimental investigation of the risk of CRISPR-induced selection of cancer gene mutants across many different cell types and lineages. I further studied whether this selection is

dependent on the Cas9/sgRNA-delivery method and/or the gene being targeted. Importantly, we asked whether other cancer driver mutations may also be selected during CRISPR-Cas9 gene editing and identified that pre-existing KRAS mutants may also be selected for during CRISPR-Cas9 editing. In summary, we established that the risk of selection for pre-existing p53 or KRAS mutations is non-negligible, thus calling for careful monitoring of patients undergoing CRISPR-Cas9-based editing for clinical therapeutics for pre-existing p53 and KRAS mutations.

In the second project, we aimed to delineate some of the molecular mechanisms that may underlie the observed differences in cancer incidences across cancer patients of different ancestries, focusing mainly on lung cancer. We found that lung tumors from African American (AA) patients exhibit higher genomic instability, homologous recombination deficiency, and aggressive molecular features such as chromothripsis. We next demonstrated that these molecular differences extend to many other cancer types. The prevalence of germline homologous recombination deficiency (HRD) is also higher in tumors from AAs, suggesting that at least some of the somatic differences observed may have genetic origins. Importantly, our findings provide a therapeutic strategy to treat tumors from AAs with high HRD, with agents such as PARP and checkpoint inhibitors, which is now further explored by our experimental collaborators.

In the third project, we developed a new computational framework to leverage single-cell RNA-seq from patients' tumors to guide optimal combination treatments that can target multiple clones in the tumor. We first showed that our predicted viability profile of multiple cancer drugs significantly correlates with their targeted pathway activity at a single-cell resolution, as one

would expect. We apply this framework to predict the response to monotherapy and combination treatment in cell lines, patient-derived-cell lines, and most importantly, in a clinical trial of multiple myeloma patients. Following these validations, we next charted the landscape of optimal combination treatments of the existing FDA-approved drugs in multiple myeloma, providing as a resource that could be used to potentially guide combination trials.

Taken together, these results demonstrate the power of multi-omics analysis of cancer data to identify potential cancer risks and a strategy to mitigate, to shed light on molecular mechanisms underlying cancer disparity in AA patients, and point to possible ways to improve their treatment, and finally, we developed a new approach to treat cancer patients based on single-cell transcriptomics of their tumors.

THREE VARIATIONS OF PRECISION MEDICINE: GENE-AWARE GENOME EDITING,
ANCESTRY-AWARE MOLECULAR DIAGNOSIS, AND CLONE-AWARE TREATMENT
PLANNING

by

Sanju Sinha

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:

Professor Steve Mount, Chair (Co-Advisor)

Dr. Eytan Ruppin (Advisor)

Dr. Brid Ryan (Co-Advisor)

Dr. Rob Patro

Professor Sridhar Hannenhalli

Professor Najib M. El-Sayed

© Copyright by
Sanju Sinha
2021

Dedicated to my parents, Sukhnandan Prasad Sinha and Munni Devi, whose love, dedication,
and sacrifices provided me the platform to reach this point in life.

Acknowledgments

Firstly, I would like to thank Eytan Ruppin, my chief advisor, for this opportunity, mentorship, and support throughout this journey. With his critical and energetic guidance, this journey has been an exciting challenge with many scientific and non-scientific lessons to keep and cherish for a long time. His care, kindness, and support extended beyond academic life throughout the highs and lows during this period and I am deeply grateful and fortunate to receive them.

Secondly, I would like to thank Brid Ryan, my co-advisor, for her invaluable mentorship, support and patiently teaching me countless fundamentals. Her confidence in me has provided me the essential mental support throughout this journey and I am indebted to her for this. I would cherish our discussions for a long time. I am extremely blessed to have such a great pair of mentors. As much as I am excited to take the next steps of my career, my heart is sad that I have to leave soon.

During my Ph.D., I have had the honor to work with many excellent scientists, including Curtis Harris, Ze'ev Ronai, Ani Deshpande, Moshe Oren, Kenneth Aldape, Max Leiserson, Silvio Gutkind, and many more, and have learned a wide and diverse set of scientific lessons from them, and I would like to thank them all.

With no less enthusiasm than before, I would like to thank my colleagues and friends at the Cancer Data Science Lab at NIH and UMD. It might be impossible to name everyone and show my gratitude for their support, however, to name a few - Sanna Madan, Fiorella Schischlik, Kun Wang, David Crawford, Ethan Iverson, and many more. Including our lunches, hikes, and happy hour, I would keep these memories close to my heart. It has been a true pleasure and a gem of an experience to work closely with Kuoyuan Cheng during this time. I have learned

numerous lessons while working closely with Alejandro Schaffer, Joo Sang Lee, and Rahul Vegesna and would also like to thank them.

Starting from day 1, the staff at UMD have been actively kind, helpful, and truly caring, including Michelle Brook, Gwen Warman, Zakiya Whatley, Hadiya Woodham & Barbara Lewis. At NIH, a shoutout and special thanks to Nadiya Nimley for her help and kindness during the last few years.

Looking back, a critical part of this journey was the companionship of two of my close friends, Anshuman Bhanja and Jurrian Van Haaren, and am very grateful and fortunate to have this. Importantly, I would like to thank my girlfriend Samprita for her support, encouragement, and unconditional love throughout the last two years. She has shared my highs with celebrations and lows with support, care, and encouragement throughout this time.

Lastly and most importantly, I would like to thank my family, especially my parents for their incomparable love, support, and guidance throughout my life. I would not have reached here without them and their sacrifices and support.

Table of Content

Chapter 1: Higher prevalence of homologous recombination deficiency in tumors from African Americans versus European Americans	1
Chapter 2: A systematic genome-wide mapping of oncogenic mutation-selection during CRISPR-Cas9 genome editing	39
Chapter 3: Designing optimal combination treatment targeting the clonal architecture of the tumor using scRNA-seq	82
Conclusion	112

Chapter 1: Higher prevalence of homologous recombination deficiency in tumors from African Americans versus European Americans

To improve our understanding of longstanding disparities in incidence and mortality in lung cancer across ancestry, we performed a systematic comparative analysis of molecular features in tumors from African Americans (AAs) and European Americans (EAs). We find that lung squamous cell carcinoma tumors from AAs exhibit higher genomic instability—the proportion of non-diploid genome—aggressive molecular features such as chromothripsis and higher homologous recombination deficiency (HRD). In The Cancer Genome Atlas, we demonstrate that high genomic instability, HRD, and chromothripsis among tumors from AAs are found across many cancer types. The prevalence of germline HRD (that is, the total number of pathogenic variants in homologous recombination genes) is higher in tumors from AAs, suggesting that the somatic differences observed have genetic ancestry origins. We also identify AA-specific copy-number-based arm-, focal- and gene-level recurrent features in lung cancer, including higher frequencies of PTEN deletion and KRAS amplification. These results highlight the importance of including under-represented populations in genomics research.

Introduction

In the United States (US), African Americans (AAs) have the highest cancer incidence and lowest survival across multiple cancer types ¹. The reasons for these persistent trends are not clear. Our current understanding of the molecular mechanisms of tumorigenesis is primarily from analyses of tumors derived from European ancestry patients, including The Cancer Genome Atlas (TCGA) where only 8.5% of samples are from AAs. This raises a question about whether there are differences in tumor evolution and molecular features by genetic ancestry. Recently, Yuan *et al.* compared somatic copy number alteration (SCNA)-based genomic instability (GI) across genetic ancestry in TCGA and found that invasive breast carcinoma (BRCA), head and neck squamous cell carcinoma (HNSC), and uterine corpus endometrial carcinoma (UCEC) tumors from AAs had significantly increased GI compared with tumors from European Americans (EAs) ². Further, recent work demonstrated that the African pan-genome contains numerous large insertions—whose total length comprises ~10% of the genome—that are not present in the current human reference genome (GRCh38) ³, which was primarily derived from a small number of individuals, primarily of European descent ⁴. Together these data highlight the need for studies specifically investigating the molecular landscape of cancer in minority and under-represented populations.

Lung cancer, the second most common cancer in the US and the leading cause of cancer-related death ⁵, has persistent disparities in both incidence and mortality. AAs have the highest lung cancer incidence and mortality rates when compared with other racial or ethnic groups ^{1,6}. These disparities persist even after considering tobacco smoking exposure, the strongest risk factor for lung cancer development ⁶.

Population-specific molecular patterns in tumor biology and cancer genomics have been reported in recent years ⁷⁻¹⁰ with limited power and coverage. Here, we systematically identified ancestry-specific genome-wide copy number features in a racially balanced (EA and AA) cohort of 222 lung tumors. Our analysis reveals higher GI and homologous-recombination deficiency (HRD) in LUSC tumors from AAs compared with EAs. This suggests an ancestry-associated disparity in deficiency of the HR-pathway, which we confirmed by finding a higher prevalence of germline HRD in AA compared with EA patients in LUSC. In the TCGA cohort, we further found the increased GI, HRD, and chromothripsis (CHTP) among AAs across multiple cancer types and pan-cancer. Further, we identify ancestry-specific arm, focal, and gene-level features in LUAD and LUSC. Our results highlight the importance of including minority and under-represented populations in cancer genomics research and may have therapeutic implications.

Results

LUSC tumors from African Americans have higher GI and HRD

We generated genome-wide copy number profiles of 222 non-small cell lung cancer tumor samples from the NCI-MD study (105 LUAD [AA=63, EA=42] and 117 LUSC [AA=63, EA=54] (Supp Table 1) using the OncoScan platform ¹¹, which provides comprehensive coverage of 50–100 kb copy number resolution in cancer genes and 300 Kb across the rest of the genome. Sample characteristics for the patients in this study are shown in Supp Table 1. Based on these copy-number alterations profiles, we first quantified GI—defined as the proportion of the genome with non-diploid copy number—for each sample. We found that LUSC tumors from AAs had significantly higher GI compared with EAs (Figure 1A-top panel;

Wilcoxon Rank-Sum (shortened to ‘Wilcoxon’ henceforth) $P < 6E-03$). In contrast, we did not find significantly higher GI in lung adenocarcinoma (LUAD) in AAs (Figure 1A-middle panel). We tested the hypothesis that higher GI across tumors from AAs is due to a higher prevalence or extent of HRD, which was previously identified as a key contributor to GI in cancer ¹². We quantified HRD in tumors using four independent measures of HRD: Loss of heterozygosity (LOH), which is the number of LOH segments ^{13,14}; telomere allelic imbalance (AIL), which is the number of regions of allelic imbalance that extend to one of the sub-telomeres but do not cross the centromere; large-scale state transitions (LST), which is the number of breakpoints between regions longer than 10 Mb after filtering out regions shorter than 3 Mb ¹³ and lastly, the sum of these three features. All four scores are scaled within the range of 0 to 1. In the NCI-MD study, we observed a strong positive correlation between GI and HRD across the whole cohort for all four features ($P < 2E-16$ for all; Spearman Rho=0.64 for LOH, 0.31 for LST, 0.44 for AIL, 0.51 for the sum), where, in AA tumors, the correlation observed is stronger than in EA tumors (Spearman Rho for AA=0.57, for EA=0.48, $P < 2.2E-16$ for both) (Supp Table 2). Next, we observed significantly higher HRD in AAs with LUSC (FDR adjusted P-value $< 2E-04$ for LOH, Figure 1B-top panel; $< 2.0E-02$ for LST; $< 3.9E-02$ for AIL; $P < 7.1E-03$ for net sum), but not LUAD, which is consistent with our GI-based findings outlined above (Figure 1B-middle panel). This suggests that HRD contributes to the ancestry-specific pattern of higher GI burden in LUSC among AAs.

To account for potential confounding factors, we performed multivariate linear regression to model separately GI and HRD in the NCI-MD cohort as a function of ancestry adjusting for tumor stage, patient age, sex, smoking status, pack-years of cigarettes and tumor purity. Here, we found AA ancestry strongly positively associated with GI and HRD in LUSC, but not LUAD,

consistent with our previous observations (LUSC: $\text{FDR} < 3\text{E-}02$ and $\text{FDR} < 5.35\text{E-}05$, respectively, LUAD: $\text{FDR} < 0.24$ and $\text{FDR} < 0.09$, respectively, Supp Table 3).

We initially determined ancestry by self-report; however, it is possible that miss-report could have confounded our results¹⁵. Therefore, we inferred ancestry in an unsupervised manner via principal component analysis (PCA) of ancestry-informative SNPs (Methods) followed by classification of the first two PCs via support vector classification (SVC), which identified two classes of ancestry. We found that inferred ancestry class is concordant with self-reported ancestry for 98.6% of subjects; four samples were potentially misclassified (Supp Table 3, Column B). We removed these samples and repeated the analyses above and found consistent results with comparable significance (higher GI and LOH-HRD in LUSC among AAs with Wilcoxon $P < 6\text{E-}03$ and $P < 2\text{E-}04$).

To validate the relationship between GI and the extent of HRD that we found in the NCI-MD cohort, we quantified GI and HRD using the four signatures described above in the TCGA cohort. Both GI and HRD were higher in tumors from AAs compared with EAs in LUSC, but the differences did not reach statistical significance (Figure 1A-B, bottom Panel). This could be due to the limited number of AA tumor samples in TCGA (29 AA compared with 346 EA), which is supported by a power analysis of TCGA samples across ancestry (Methods).

Lung tumors from African Americans have more frequent complex structural variants

The observed deficiencies in DNA damage repair related with GI in LUSC prompted us to chart the landscape of complex structural variants recently reported to be related with HRD¹⁶. We studied chromothripsis (CHTP), which was first described as a catastrophic event that leads to chromosome shattering and tens to hundreds of simultaneously acquired oscillatory copy number

aberrations on one chromosome^{17,18}. Therefore, we represented CHTP as a binary variable indicating presence/absence. Using the classical definition, i.e., many oscillatory copy number events clustered on a chromosome (Methods)¹⁹, tumor samples with CHTP had significantly higher HRD than samples without CHTP (Wilcoxon $P < 9E-04$) in the NCI-MD lung cancer cohort (Supp Table 2). Further, we observed higher frequency of CHTP in tumors from AAs compared with EAs in LUSC (Figure 1C-Top panel, $P < 0.12$, OR=1.24) and in LUAD, but to a weaker extent (Figure 1C-Middle panel, $P < 0.49$, OR=1.15). These patterns are consistent when adjusted for age, sex stage, smoking status and pack-years of cigarettes (multivariate regression P for ancestry $< 2.8E-03$, Supp Table 3). The same result held qualitatively when an alternative quantification of CHTP, defined by the allowance for two oscillation states in the affected region, was used (Methods) (Supp Table 2- Column AD). Next, we quantified CHTP in the TCGA-LUSC cohort and observed a consistent pattern of higher frequency in tumors from AA (Figure 1C-bottom panel, $P < 0.12$, OR=1.45). We further analyzed the chromosome frequency distribution of CHTP, which varied by histological subtype and ancestry.

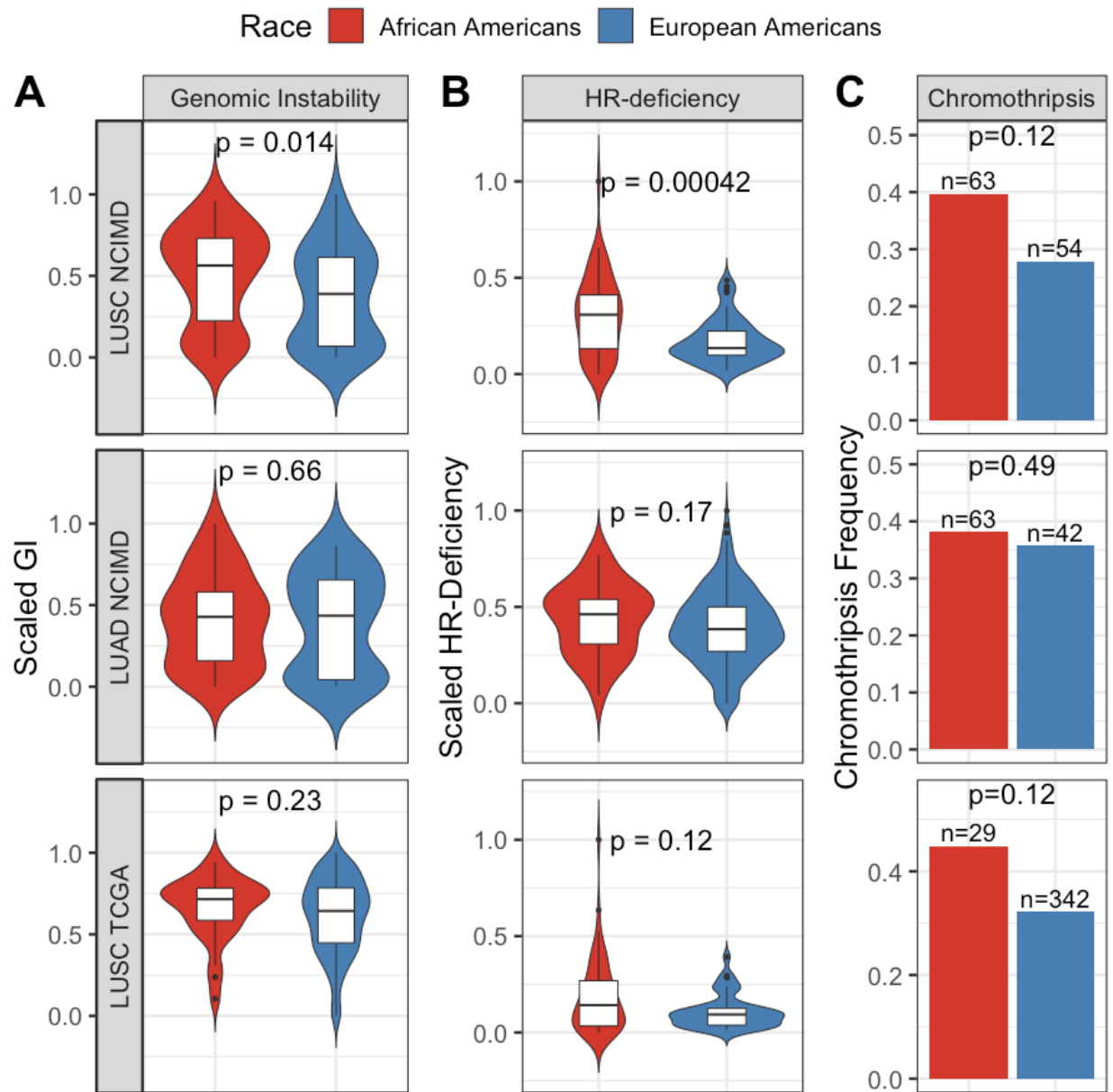


Figure 1: Differences in GI, HRD and chromothripsis across AA and EA lung cancer patients from the NCI-MD and TCGA cohort. (A) Genomic instability, (B) homologous recombination deficiency and (C) chromothripsis are quantified and presented stratified by genetic ancestry in LUSC (top, $n=105$ patients (AA=63, EA=42)) and LUAD (middle, and $n=117$ patients (AA=63, EA=54)) samples from the NCI-MD cohort and LUSC from the TCGA

cohort (bottom, and n=375 patients (AA=29, EA=346)). *LUAD* denotes lung adenocarcinoma and *LUSC* denotes lung squamous cell carcinoma. Significance for comparison of medians in A) and B) is calculated via one-sided Wilcoxon rank-sum tests and significance for comparison of frequency in C) is calculated via one-sided Fisher's exact test. The violin plots in A) and B) show the data distribution where the center line denotes the median, the box indicating the interquartile range and the black line represents the rest of the distribution, except for points that are determined to be "outliers" using a method that is a function of the interquartile range, as in box plots.

The landscape of arm- and focal-level SNCAs in AA and EA lung cancer

To identify SCNA-based ancestry-specific features in detail, we examined population-specific SCNA profiles in lung cancer for chromosome arm- and focal- level (shorter than half a chromosome arm) events in the NCI-MD study where statistical power for both populations was available. Further support for key observations was demonstrated in TCGA. Recurrent arm-level and focal-level SCNA events were identified for both populations separately using GISTIC²⁰ (Methods, FDR<0.1) and used to map genome-wide SCNA across histology and ancestry (Figure 2, Supp Tables 4-5).

For each chromosome arm, the alteration frequency and the recurrence significance by ancestry for both amplifications and deletions were plotted for patients in the NCI-MD cohort (Figure 2). We identified known cancer-specific arm-level SCNA events, including amplification of 3q and 5p and deletion of 3p²¹, in both populations (Supp Tables 4-5). Similarly, 19p deletion, a molecular signature of LUAD, was recurrent in EAs and AAs at similar frequencies of ~45% (Figure 2, Supp Table 5). Recurrent population-specific arm-level SCNA differences

were observed, including 4p and 4q arm level deletions in LUSC and 7p and 7q amplifications in LUAD, both occurred at higher frequency in AAs compared with EAs. These observations were replicated in TCGA (Figure 2).

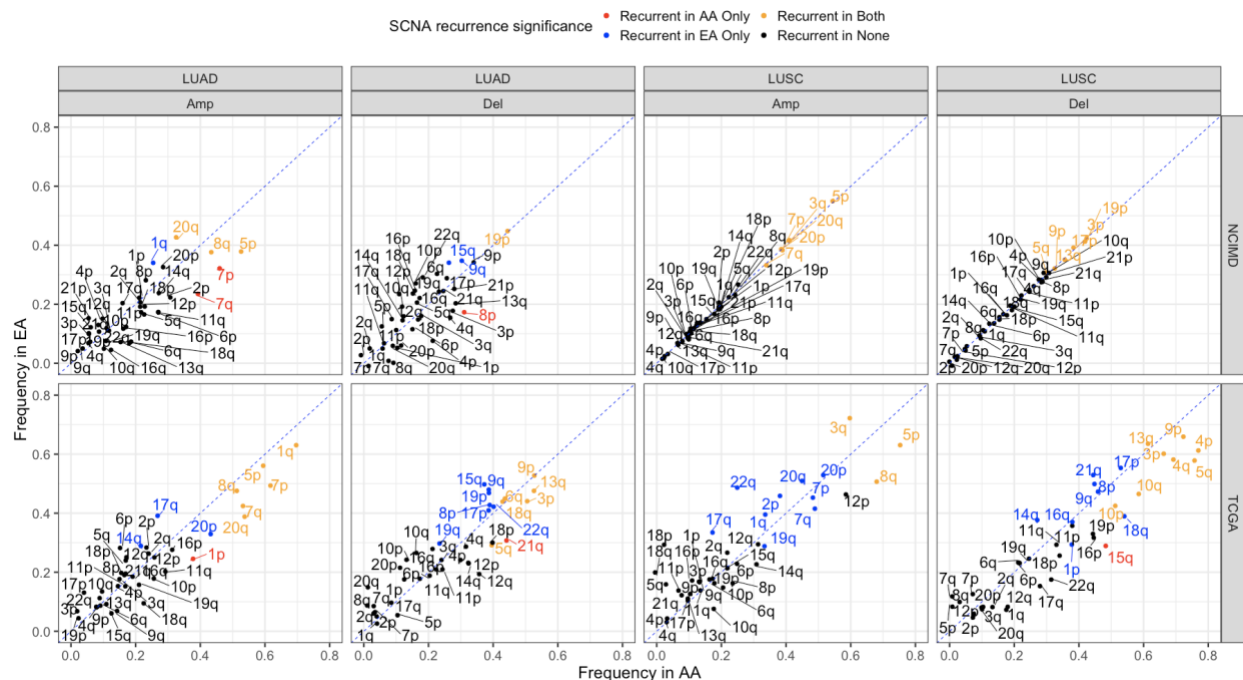


Figure 2: Characterization of arm-level SCNA events across EA and AA lung cancer in the NCI-MD cohort. Frequency distribution of aberrant SCNA events on autosomal chromosome arms in LUAD and LUSC for the NCI-MD and TCGA cohorts (LUSC n=375 patients [AA=29, EA=346], LUAD n= 432 patients [AA=51, EA=381]). The diagonal dashed line represents equal AA and EA frequencies, with points falling away from this line indicating chromosome arms with alteration frequency differences between populations. A color code is provided to denote population-specific recurrent SCNA events with statistical significance. Del=deletion, amp=amplification. Statistical significance of recurrence was computed via GISTIC, which provides arm-level FDR-corrected significance with a threshold of 0.1.

To visualize genome-wide focal-level SCNA events across populations, including co-occurrence and mutual-exclusivity, we created an SCNA map showing genome-wide SCNA frequency

distributions for both LUSC and LUAD in the NCI-MD cohort (Figure 3-left panel). The overlaps in recurrent focal regions among EAs and AAs were 59% and 70% for LUAD and LUSC, respectively (Figure 3-left panel). Further, we observed population-specific patterns of co-occurring and mutually exclusive SCNA events (Figure 3-left panel). To identify potential novel AA-specific copy number-driven focal-level regions, we selected high-confidence recurrent focal-level regions from GISTIC that met the following criteria; 1) alteration frequency greater than 5% in AAs, 2) frequency at least two times higher in AAs than EAs 3) recurrent only in AAs and 4) no recurrent peak of the same type (amplification or deletion) was present in EAs within the region or an extended additional 10% on both sides of the region length. We identified eight potential AA-specific potential driver regions. The top hit ranked by significance is a 22q11.23 deletion in LUSC (Figure 3-right panel) with a frequency of 27% in AAs and 13% in EAs. Following a previous study ²²⁻²⁴, we tested whether this deletion event is somatic or germline by profiling matched-normal tissue samples with genome-wide copy number; we observed that 2/5 normal samples from AAs also have a deletion of 22q11.23, suggesting that this event could be germline (Supp Table 6). This 22q11.23 region deleted in LUSC is disjointed from the nearby region on 22q11.21 that is hemizygously deleted in DiGeorge syndrome ²³⁻²⁴. The region with the second-highest fold change in alteration frequency in LUSC, 12p12.1 (Figure 3-right panel), is a short region including *KRAS* and is discussed in detail in the next section. Thirdly, common to both LUAD and LUSC, the 20p12.1 region is deleted >4 times as often in AAs compared to EAs. This region includes the genes *FLTR3* and *MACROD2*. We also identified several SCNA events previously linked with AA ancestry in cancer and assessed the relationship between copy number and gene expression (Supp Tables 7-10). We observed an AA-specific amplification of the oncogene *KAT6A* in LUAD, which was previously

observed in ²⁴. We also identified a recurrent deletion of 4q35.2 extending to the telomere in LUSC that includes *FBXW7*, previously shown to be deleted in colorectal cancer and triple-negative breast cancer among AAs ^{25,26} (Supp Table 7). In LUAD, a region in near 8q24 was significantly recurrently amplified in AAs only (frequency=33% and 18% in AAs and EAs, respectively). Within a sub-region, 8q24.21, the *PVT1* copy number profile was significantly associated with expression ($P < 7E-03$), while in 8q24.3, *HSF1*, *DGAT1*, and *BOP1* copy number were also significantly associated with gene expression ($P < 7E-03$) (Supp Tables 8 & 10).

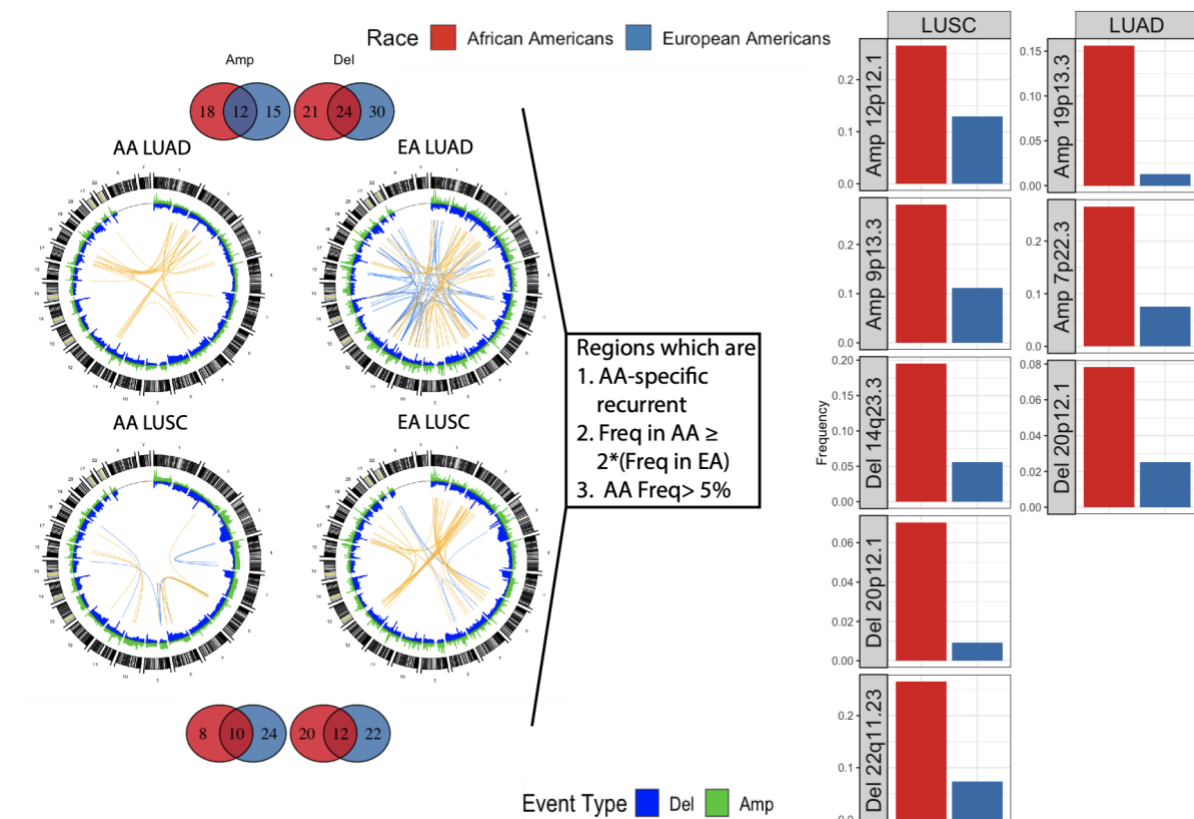


Figure 3: Global SNCA map across EA and AA lung cancer in the NCI-MD cohort.

Segmental deletions and amplifications are shown in blue and green, respectively in the left panel circos plot. In this plot, the top 50 ($|\text{Pearson-Rho}| > 0.50$) highly positively (co-occurring) and negatively (mutually exclusive) correlated copy number segment pairs are connected with

yellow and blue arcs, respectively. The overlap and unique recurrent regions between AAs and EAs in LUSC and LUAD are shown via Venn diagrams at the top and bottom. Steps provided in the central box are used to identify SNCA-driven AA-specific potential-driver regions, where the list of regions passing these steps are provided with corresponding frequency in AAs and EAs on the right via bar plots for LUAD and LUSC. Recurrence significance for each focal region was computed via GISTIC in AAs and EAs separately with an FDR-corrected significance threshold of 0.1.

The landscape of driver genes SCNAs in AA and EA lung cancer

We analyzed the recurrence and alteration frequency of known lung cancer driver genes mined from the cancer gene census of COSMIC (Figure 4A). We identified population-specific SCNA patterns of drivers (Figure 4A) significantly correlated with gene expression (Figure 4B and Figures S3-4). In LUSC, one of the key cancer driver genes, *KRAS*, is amplified in both populations but is significantly recurrent (FDR<0.1, Methods) and has a higher frequency in AAs (*KRAS* amp frequency: 23% in EAs compared to 51% in AAs, Figure 4). Similarly, *PTEN* deletion is significantly recurrent and more frequent in AAs (*PTEN* del frequency: 32% in EAs compared to 53% in AAs, Figure 4). Another key driver, *CDKN2A*, was recurrently deleted in both populations, but the frequency was 35% in AAs compared with 64% in EAs (Figures 4A). These three population-specific patterns in frequency were also observed in TCGA.



Figure 4: Landscape of SCNA of lung cancer drivers across EA and AA lung cancer in the NCI-MD cohort. (A) Amplification and deletion frequency of lung cancer driver genes across population and histology. Recurrence significance for each gene was computed via GISTIC in AAs and EAs separately with an FDR-corrected significance threshold of 0.1. The diagonal dashed line denotes the null line with points falling away from this line indicating chromosome arms with alteration frequency differences across populations. A color code at the top of panel 4A is provided to denote gene-level population-specific statistically significant recurrent SCNA events, where a gene name being in black implies no statistically significance SCNA recurrence in either population. Del=deletion, amp=amplification. (B) Effect of copy number changes on expression profile (n=91 patients) of drivers with population-specific patterns. Here, at the top of each panel, we have provided the corresponding gene name, with the two-sided Spearman significance (P) and Rho (ρ) of the mRNA expression of the gene and its SCNA profile. Only

genes significantly correlated with expression are plotted ($P < 0.01$ & Spearman $Rho > 0.2$). Here, the centerline denotes the median, the box indicating the interquartile range, and the black line represents the rest of the distribution, except for points that are determined to be “outliers” using a method that is a function of the interquartile range, as in box plots.

A pan-cancer survey of GI, HRD, and chromothripsis in AA vs EA tumors

To determine whether a higher prevalence of aggressive molecular features, including GI, HRD, and CHTP, extends to other cancer types, we mined TCGA SCNA profiles of 6,492 tumor samples with available self-reported ancestry from AAs and EAs originating from 23 tumor types (Supp Tables 11 & 12). Consistent with previous observations³, we initially observed an overall significantly higher GI burden in AA tumors (pan-cancer Wilcoxon $P < 6.9E-07$, Figure 5A). These differences were most significant in breast (BRCA), head and neck (HNSC), stomach adenocarcinoma (STAD), cervical squamous cell carcinoma, and endocervical adenocarcinoma (CESC) cancers, with a general trend towards higher GI burden in 17 out of the 23 cancer types. We repeated this analysis separately for SNCA-loss and -gain-based GI and observed a consistent pattern (Methods).

We quantified HRD using the four measures previously used, i.e., LOH, TIL, LST and normalized sum of the three and observed a strong correlation between GI and HRD in pan-cancer ($P < 2E-16$ for all; Spearman $Rho = 0.56$, Spearman $Rho = 0.47$ for LST, 0.60 for AIL, 0.58 for sum) and cancer type-specific analyses (Supp Table 13), where, in AA tumors, the correlation observed is stronger than in EA tumors for both pan-cancer (LOH-based measure: Rho for AA=0.57, for EA=0.48, $P < 2.2E-16$ for both; AIL-based measure: Rho for AA=0.66, for EA=0.60, $P < 2.2E-16$ for both; LST-based measure: Rho for AA=0.51, for EA=0.47, $P < 2.2E-16$

for both) and cancer type-specific analyses (Supp Table 13). Moreover, HRD is significantly higher in AAs in pan-cancer for all four measures (Wilcoxon $P < 1.5E-02$; $P < 7.7E-02$ for LST; $P < 2.2E-02$ for AIL; $P < 1.9E-02$ for sum). This further suggests that HRD contributes to the ancestry-specific pattern of higher GI burden in AAs across cancer types.

When analyzed by specific cancer type, we find that BRCA and HNSC have significantly higher HRD across all four measures in AAs compared with EAs (Table S12). A trend towards increased HRD among AAs was observed in 11 out of 17 cancer types where increased GI was also observed. The remaining six cancer types had an inverse trend, including KIRP and KIRC, which have significantly lower GI and HRD. We confirmed these results by quantifying HRD using a somatic mutation profile-based signature²⁷, i.e., mutational signature (mutSig) 3. This signature is typified by a C>G/A transversion and is strongly associated with HRD²⁷⁻²⁹. We leveraged the mSignatureDB database where mutation signatures are profiled²⁷ on 7,042 tumors from 30 different cancer types and found the mutSig 3 contributions to be higher in tumors from AAs compared with EAs in pan-cancer (Wilcoxon $P < 1E-03$). Testing each cancer type specifically for a higher mutSig 3 in AAs, we found that BRCA and HNSC have a higher prevalence of this HRD-related signature, which is consistent with the SNCA hallmarks-based quantification of HRD described above (Wilcoxon, $P < 0.01$ and $P < 0.10$, respectively). We additionally performed a multivariate regression modeling GI and HRD in pan-cancer as a function of ancestry adjusting for stage, sex, age and smoking status in TCGA samples, and found AA ancestry strongly positively associated with these genomic features, i.e. higher GI (FDR $< 2.2E-07$) and HRD (FDR $< 4.5E-06$ for LOH, $< 7.8E-07$ for AIL, $< 3.8E-05$ for LST, $< 2E-01$ for mutSigs3) (Supp Table 3).

Similar to the NCI-MD cohort, we tested for possible confounding by mislabeled self-reported race. We accessed genotype information of 906,600 SNPs in matched PBMCs that were downloaded from the controlled access part of TCGA (Methods) and inferred unsupervised ancestry (Methods). The overall concordance of our computed inferred ancestry with self-reported ancestry is high (94.7%). Using this inferred ancestry, we removed the possibly misclassified samples and repeated the above analysis, with consistent significant results.

Next, we quantified CHTP in TCGA samples and observed that tumor samples with CHTP have significantly higher HRD than samples without CHTP (Wilcoxon $P < 3.2E-10$, for all five HRD markers) in both pan-cancer and cancer type specifically. Consistently, we observed a higher frequency of CHTP in tumors from AAs compared with EAs in pan-cancer samples (Figure 5C, Fisher's one-sided test $P < 0.028$, odds ratio (OR)=1.25) and in LUSC samples from TCGA (Figure 5C, $P < 0.11$, OR=1.4). These patterns were consistent when adjusted for age, sex, and stage across both cohorts (multivariate regression P for ancestry $< 2.8E-03$) and further when another CHTP definition was used (Methods). Similar to the NCI-MD cohort, we observed chromosome enrichment of CHTP on chromosome 12 among AAs in LUSC (Extended Data Figure 2).

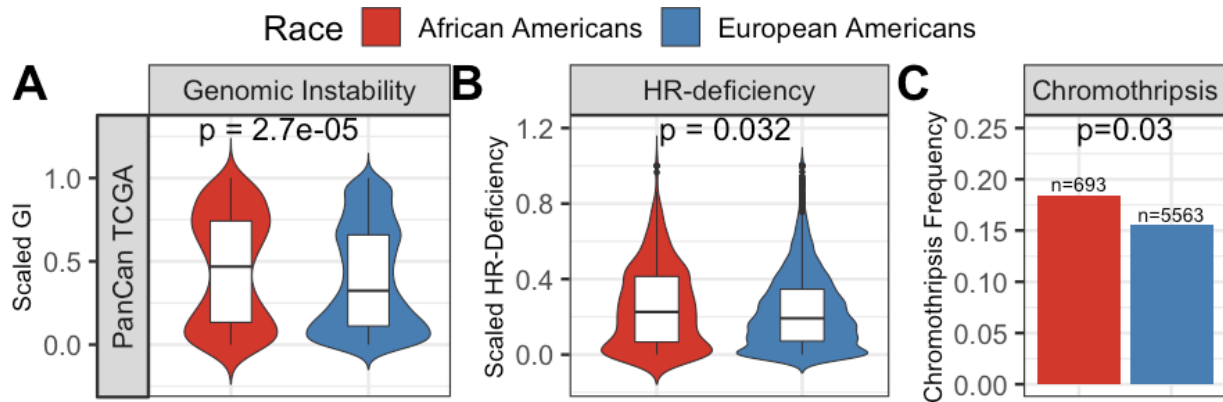


Figure 5: Landscape of GI, HRD, and chromothripsis across AA and EA pan-cancer in the TCGA cohort. (A) Genomic instability, (B) homologous recombination (HR) deficiency, and (C) chromothripsis is quantified and provided across genetic ancestry in pan-cancer TCGA samples (n=6,256 patients [AA=692, EA=5,563]). Similar to Figure 1, significance for comparison of medians in A) and B) was calculated via one-sided Wilcoxon rank-sum tests, and significance for comparison of frequency in C) was calculated via one-sided Fisher’s exact test. The violin plot shows the data distribution where the centerline denotes the median, the box indicating the interquartile range, and the black line represents the rest of the distribution, except for points that are determined to be “outliers” using a method that is a function of the interquartile range, as in box plots.

AA tumors have a higher germline prevalence of HRD

Given the higher prevalence of HRD in AA tumors across LUSC and pan-cancer, we asked whether the increase of HRD in tumors could be driven by germline factors? We accessed the TCGA database of germline pathogenic variants across 10,389 adult-tumors³⁰. This study³⁰ performed whole-exome sequencing on PBMCs and then cataloged pathogenic variants (Methods). Using this dataset, we first counted the total number of pathogenic variants in HR-

genes (Supp Table 14) in each patient and defined it as *germline* HRD. Next, we asked whether AA patients have a higher extent of germline HRD than EA patients. In TCGA pan-cancer and LUSC, but not LUAD, we found that AAs had significantly higher germline HRD than EAs (Figure 6-left panel, OR=1.2; $P<0.02$ for pan-cancer; Figure 6-right panel, OR=6; $P<8E-04$ for LUSC; Extended Data Figure 10, $P<0.23$ for LUAD). Repeating this analysis in LUSC patients for individual genes of the HR pathway, we found predicted pathogenic variants in canonical HR-pathway genes *BRCA1*, *BRCA2*, and *POLD1* to be enriched in AAs (Supp Table 14) (hypergeometric $P<0.15$, 0.01, 0.08, respectively). Similarly, in pan-cancer we found predicted pathogenic variants of *BARD1*, *FANCM*, *BRIP1*, *PALB2*, *POLD1*, and *BRCA2* to be more enriched in AA patients ($P<0.06$, 0.12, 0.12, 0.19, 0.2, 0.25, respectively). Since some of these genes are mutated in hereditary predisposition syndromes it is possible that AAs in TCGA have a higher incidence of such syndromes. However, the known syndromes do not necessarily match the observed LUSC cancer type. *BRCA2* mutations most commonly predispose to breast and ovarian cancers, although there is some evidence of association with lung cancer³¹. Mutations in *POLD1* have been associated with colorectal cancer³², but not lung cancer, to our knowledge. We also found *BLM* and *RECQL* predicted pathogenic variants to be more enriched in EA patients ($P<0.06$, 0.22).

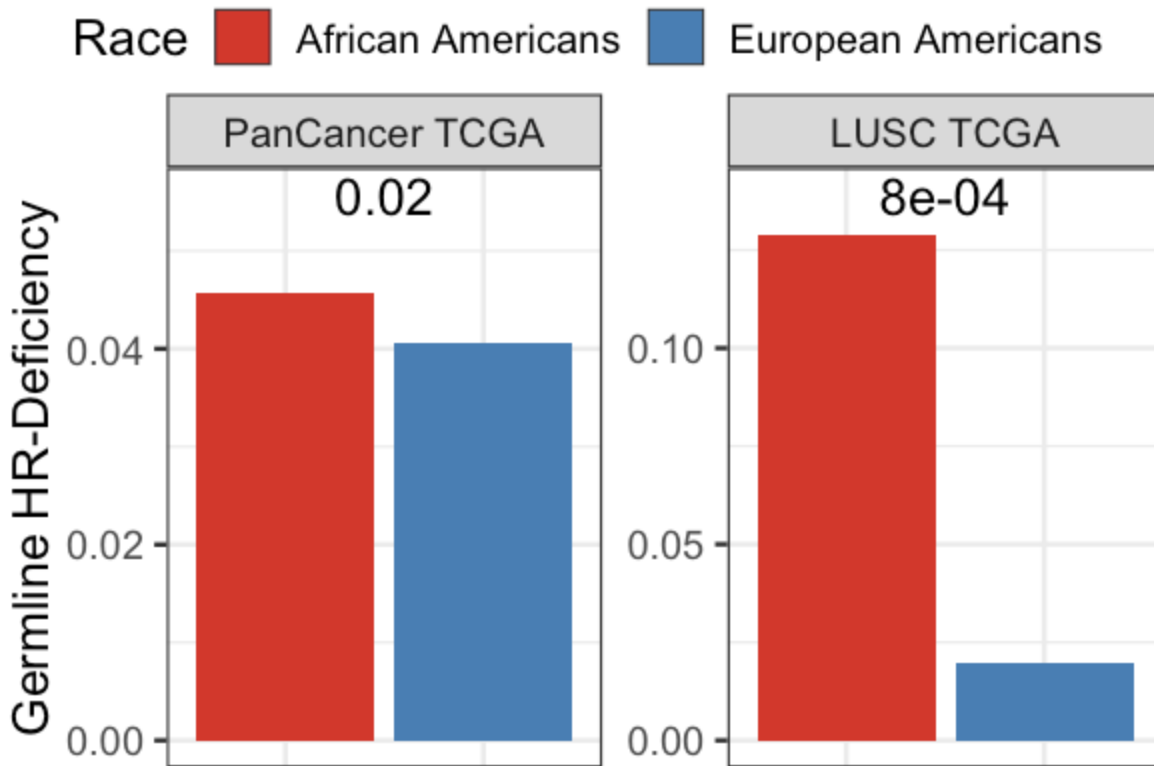


Figure 6: Landscape of germline HRD across EA and AA pan-cancer and LUSC TCGA cohort. Prevalence of germline HRD in AAs and EAs calculated using total frequency of germline pathogenic variants in HR-pathways genes in pan-cancer (left, total n=8,920 patients [AA=919, EA=8,001]) and LUSC (right, total n=382 patients [AA=31, EA=351]). Significance for comparison of the frequency of germline HRD was calculated via one-sided Fisher's exact test where exact p-value values are provided.

Discussion

Here, we mapped molecular features of tumors from EAs and AAs across many cancer types, with greater depth and power in lung cancer. We observed that consistent with previous reports ², GI is higher in AAs across multiple cancer types. This higher GI is unlikely to be related to the

recently identified unmapped 10% of the genome that is found in populations of African ancestry³ as we found both copy number gain- and copy number loss-based GI to be higher in AAs. We hypothesized, and confirmed, that this higher GI is likely due to a higher prevalence of HRD in tumors from AAs. We also identified a significantly higher prevalence of mutational signature 3—closely associated with HRD²⁷⁻²⁹—among a wide range of tumors from AAs (Extended Data Figure 8). We further show that tumors from AAs have a higher frequency of aggressive molecular features, including structural variants. HRD was not uniformly higher among AAs in some cancers, including KIRP and KIRC, where HRD was significantly lower.

Higher SNCA-based GI and HRD in tumors from AAs raises the question of whether underlying defective DNA repair mechanisms could drive this observation. While HRD has been linked with germline and somatic mutations in *BRCA1/2*³³, no striking differences in the somatic mutation frequencies of these genes have been demonstrated in cancer between EAs and AAs^{2,34}. To investigate whether the increased HRD could be driven by a germline event, we analyzed germline pathogenic variants³⁰ and identified a higher proportion of HRD-related pathogenic variants among AAs compared with EAs, suggesting that GI/HRD events and tumor evolution could be shaped by these features. The observation that several cancer types occur at an earlier age among AAs³⁵ and evidence that germline pathogenic events are associated with early-onset disease³⁰ are consistent with these data.

Higher HRD in LUSC and many other cancer types suggests that these tumors could respond to PARP inhibitors and that perhaps, AAs may be more likely to respond. Most trials do not report and/or are not powered to compare differences in response by ancestry group. PARP inhibitors are not commonly used in lung cancer treatment, though in combination with chemotherapy they have shown promising efficacy in both cell lines³⁶ and a clinical trial³⁷. In the latter, the benefit

from the combination treatment was primarily restricted to LUSC tumors. Further, a recent retrospective analysis of clinical trial data found that response to platinum compounds and survival was significantly better in patients with hallmarks of HRD ³⁸. Thus, future preclinical and clinical studies could include biomarkers of HRD either in the study design or as a covariate in the data analysis.

We next identified multiple ancestry-specific chromosome alterations with unknown relevance, including chromosomes 7p and 7q (AA frequency twice than EA). We also observed ancestry-specific patterns of co-occurrence and mutually exclusive events and recurrent focal region alterations. Further, only one out of eight potential AA-specific driver regions identified in this study have previously known driver genes (i.e., *KRAS*). Next, we found AA-specific recurrent alterations previously linked with ancestry disparities in other cancer types ³⁹⁻⁴¹, including focal deletion of 4q35.2 comprising *FBXW7*, and amplification of oncogene *KAT6A* ⁴² (18% in AA vs. 0% in EA).

In summary, we have identified population differences in molecular features, including GI, HRD, and CHTP. As these features are related to therapy response ^{13,43,44}, our findings could have therapeutic implications. We also find higher GI and HRD in LUSC among African Americans and highlight some granular differences at the SNCA level in canonical lung cancer genes, such as *CDKN2A*, *KRAS*, and *PTEN*. As our study used the same platform to compare SCNA events across EAs and AAs, it largely removes the possibility that technical artifacts could confound our observations. Defining these differences in both genome-wide and more focal regions highlights distinct differences in lung tumor biology between AAs and EAs and supports recent work showing that inherited variants and thereby, genetic ancestry, can shape

tumor evolution at a molecular level and influence the somatic nature of a tumor ⁴⁵. Finally, our work highlights the importance of including under-represented populations in balanced genomic studies of molecular patterns and cancer evolution.

Acknowledgments

We sincerely thank C. Harris for many insightful discussions. SS gratefully acknowledges the support of the NCI-UMD Cancer Research Training Fellowship. This research was supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute.

References

1. DeSantis, C. E., Miller, K. D., Goding Sauer, A., Jemal, A. & Siegel, R. L. Cancer statistics for African Americans, 2019. *CA Cancer J Clin*, doi:10.3322/caac.21555 (2019).
2. Yuan, J. et al. Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. *Cancer cell* 34, 549-560 e549, doi:10.1016/j.ccell.2018.08.019 (2018).
3. Sherman, R. M. et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet*, doi:10.1038/s41588-018-0273-y (2018).
4. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* 328, 710-722, doi:10.1126/science.1188021 (2010).
5. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA Cancer J Clin* 69, 7-34, doi:10.3322/caac.21551 (2019).
6. Ryan, B. M. Lung cancer health disparities. *Carcinogenesis* 39, 741-751, doi:10.1093/carcin/bgy047 (2018).

7. Mitchell, K. A., Zingone, A., Toulabi, L., Boeckelman, J. & Ryan, B. M. Comparative Transcriptome Profiling Reveals Coding and Noncoding RNA Differences in NSCLC from African Americans and European Americans. *Clin Cancer Res* 23, 7412-7425, doi:10.1158/1078-0432.CCR-17-0527 (2017).
8. Wallace, T. A. et al. Tumor immunobiological differences in prostate cancer between African-American and European-American men. *Cancer Res* 68, 927-936, doi:10.1158/0008-5472.CAN-07-2608 (2008).
9. Guda, K. et al. Novel recurrently mutated genes in African American colon cancers. *Proc Natl Acad Sci U S A* 112, 1149-1154, doi:10.1073/pnas.1417064112 (2015).
10. Chaisaingmongkol, J. et al. Common Molecular Subtypes Among Asian Hepatocellular Carcinoma and Cholangiocarcinoma. *Cancer Cell* 32, 57-70 e53, doi:10.1016/j.ccell.2017.05.009 (2017).
11. Foster, J. M. et al. Cross-laboratory validation of the OncoScan(R) FFPE Assay, a multiplex tool for whole genome tumour profiling. *BMC Med Genomics* 8, 5, doi:10.1186/s12920-015-0079-z (2015).
12. Knijnenburg, T. A. et al. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep* 23, 239-254 e236, doi:10.1016/j.celrep.2018.03.076 (2018).
13. Telli, M. L. et al. Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer. *Clin Cancer Res* 22, 3764-3773, doi:10.1158/1078-0432.CCR-15-2477 (2016).
14. Swisher, E. M. et al. Rucaparib in relapsed, platinum-sensitive high-grade ovarian carcinoma (ARIEL2 Part 1): an international, multicentre, open-label, phase 2 trial. *Lancet Oncol* 18, 75-87, doi:10.1016/S1470-2045(16)30559-9 (2017).

15. Jin, Y., Schaffer, A. A., Feolo, M., Holmes, J. B. & Kattman, B. L. GRAF-pop: A Fast Distance-Based Method To Infer Subject Ancestry from Multiple Genotype Datasets Without Principal Components Analysis. *G3 (Bethesda)* 9, 2447-2461, doi:10.1534/g3.118.200925 (2019).
16. Ratnaparkhe, M. et al. Defective DNA damage repair leads to frequent catastrophic genomic events in murine and human tumors. *Nat Commun* 9, 4760, doi:10.1038/s41467-018-06925-4 (2018).
17. Zhang, C. Z. et al. Chromothripsis from DNA damage in micronuclei. *Nature* 522, 179-+, doi:10.1038/nature14493 (2015).
18. Zhang, C. Z., Leibowitz, M. L. & Pellman, D. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Gene Dev* 27, 2513-2530, doi:10.1101/gad.229559.113 (2013).
19. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144, 27-40, doi:10.1016/j.cell.2010.11.055 (2011).
20. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 12, R41, doi:10.1186/gb-2011-12-4-r41 (2011).
21. Taylor, A. M. et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer cell* 33, 676-689 e673, doi:10.1016/j.ccell.2018.03.007 (2018).
22. Polimanti, R. et al. Haplotype differences for copy number variants in the 22q11.23 region among human populations: a pigmentation-based model for selective pressure. *Eur J Hum Genet* 23, 116-123, doi:10.1038/ejhg.2014.47 (2015).
23. Burnside, R. D. 22q11.21 Deletion Syndromes: A Review of Proximal, Central, and Distal Deletions and Their Associated Features. *Cytogenet Genome Res* 146, 89-99, doi:10.1159/000438708 (2015).
24. Baell, J. B. et al. Inhibitors of histone acetyltransferases KAT6A/B induce senescence and arrest tumour growth. *Nature* 560, 253-257, doi:10.1038/s41586-018-0387-5 (2018).

25. Brim, H. et al. Genomic aberrations in an African American colorectal cancer cohort reveals a MSI-specific profile and chromosome X amplification in male patients. *PLoS One* 7, e40392, doi:10.1371/journal.pone.0040392 (2012).
26. Craig, D. W. et al. Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. *Mol Cancer Ther* 12, 104-116, doi:10.1158/1535-7163.MCT-12-0781 (2013).
27. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* 500, 415-421, doi:10.1038/nature12477 (2013).
28. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 534, 47-54, doi:10.1038/nature17676 (2016).
29. Ma, J., Setton, J., Lee, N. Y., Riaz, N. & Powell, S. N. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat Commun* 9, 3292, doi:10.1038/s41467-018-05228-y (2018).
30. Huang, K. L. et al. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* 173, 355-370 e314, doi:10.1016/j.cell.2018.03.039 (2018).
31. Wang, Y. et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet* 46, 736-741, doi:10.1038/ng.3002 (2014).
32. Palles, C. et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature genetics* 45, 136-144, doi:10.1038/ng.2503 (2013).
33. Timms, K. M. et al. Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. *Breast Cancer Res* 16, 475, doi:10.1186/s13058-014-0475-x (2014).
34. Campbell, J. D. et al. Comparison of Prevalence and Types of Mutations in Lung Cancers Among Black and White Populations. *JAMA Oncol* 3, 801-809, doi:10.1001/jamaoncol.2016.6108 (2017).

35. Robbins, H. A., Engels, E. A., Pfeiffer, R. M. & Shiels, M. S. Age at cancer diagnosis for blacks compared with whites in the United States. *J Natl Cancer Inst* 107, doi:10.1093/jnci/dju489 (2015).
36. Jiang, Y. et al. PARP inhibitors synergize with gemcitabine by potentiating DNA damage in non-small-cell lung cancer. *Int J Cancer* 144, 1092-1103, doi:10.1002/ijc.31770 (2019).
37. Ramalingam, S. S. et al. Randomized, Placebo-Controlled, Phase II Study of Veliparib in Combination with Carboplatin and Paclitaxel for Advanced/Metastatic Non-Small Cell Lung Cancer. *Clin Cancer Res* 23, 1937-1944, doi:10.1158/1078-0432.CCR-15-3069 (2017).
38. Kadouri, L. et al. Homologous recombination in lung cancer, germline and somatic mutations, clinical and phenotype characterization. *Lung Cancer* 137, 48-51, doi:10.1016/j.lungcan.2019.09.008 (2019).
39. Yeh, C. H., Bellon, M. & Nicot, C. FBXW7: a critical tumor suppressor of human cancers. *Mol Cancer* 17, 115, doi:10.1186/s12943-018-0857-2 (2018).
40. Zhang, Q. et al. FBXW7 Facilitates Nonhomologous End-Joining via K63-Linked Polyubiquitylation of XRCC4. *Mol Cell* 61, 419-433, doi:10.1016/j.molcel.2015.12.010 (2016).
41. Yumimoto, K. et al. F-box protein FBXW7 inhibits cancer metastasis in a non-cell-autonomous manner. *J Clin Invest* 125, 621-635, doi:10.1172/JCI78782 (2015).
42. Kytola, V. et al. Mutational Landscapes of Smoking-Related Cancers in Caucasians and African Americans: Precision Oncology Perspectives at Wake Forest Baptist Comprehensive Cancer Center. *Theranostics* 7, 2914-2923, doi:10.7150/thno.20355 (2017).
43. Farmer, H. et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* 434, 917-921, doi:10.1038/nature03445 (2005).
44. McCabe, N. et al. Deficiency in the repair of DNA damage by homologous recombination and sensitivity to poly(ADP-ribose) polymerase inhibition. *Cancer Res* 66, 8109-8115, doi:10.1158/0008-5472.CAN-06-0140 (2006).

45. Carter, H. et al. Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discov*, doi:10.1158/2159-8290.CD-16-1045 (2017).

Methods

Statistics & reproducibility

While generating genome-wide copy number profiles of NCI-MD via OncoScan, two aliquots from the same sample were used to test the reproducibility of the assay for three samples by the company (available on reasonable request). In the NCI-MD study design, no statistical method was used to predetermine sample size. In the additional cohort, TCGA, we mined copy number profile of samples publicly available and excluded cancer types with less than five tumor samples with AA ancestry to provide a minimum statistical power. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment though samples were run on the OncoScan assay in a blinded manner.

The processed tables to reproduce our figures and conclusions are provided (Data availability). In this work, we used non-parametric tests using R, including Wilcoxon rank-sum tests to compare the difference in medians, Fisher's tests to compare frequency, and Spearman's correlation, with an FDR-corrected P threshold of <0.1 indicating statistical significance. Wherever GISTIC was used, the FDR-corrected significance threshold of <0.1 was applied to identify significantly recurrent regions. While identifying chromothripsis, the distance between events on a chromosome is compared to the overall distance between events in the samples to identify clustered events using an FDR-corrected P threshold of <0.1 .

Samples preparation and processing

Sample characteristics

Patients living in the Baltimore metropolitan area with histologically confirmed cases of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) were prospectively recruited to the ongoing NCI-MD Case-Control Study ⁴⁶. Institutional Review Boards at seven participating Baltimore hospitals and the NCI approved the study with written informed consent obtained from all patients. All samples were collected from an NCI IRB-approved study. We conducted a retrospective study of eligible participants that self-reported as AA or EA, with non-Hispanic ethnicity. Additional clinical and sociodemographic data for each patient were obtained from medical records and pathology reports. Macro-dissected primary lung tumor tissues were obtained from patients directly after surgical removal. Samples were placed in collection tubes, flash-frozen, and stored at -80°C until the OncoScan analyses were performed. Sample characteristics for the patients in which tumor DNAs were extracted can be found in Supp Table 1 (n=142 AA, 108 EA).

DNA extraction

DNA was extracted from fresh frozen micro-dissected primary lung tumor tissues using the Qiagen DNeasy Blood and Tissue kit spin-column procedure according to the manufacturer's

protocol (Qiagen, Valencia, CA). Isolated primary lung tumor DNAs were initially quantified using a DS-11 spectrophotometer (DeNovix, Wilmington, DE). Subsequent Qubit fluorometer analyses were performed to assess DNA integrity and ensure the presence of intact double-stranded DNA of all samples (Invitrogen, Carlsbad, CA). DNA with an A260/A280 ratio between 1.8 and 2.0, a minimum concentration of 12 ng/ μ L, and a total concentration of 80 ng was used for further analysis.

Preprocessing of raw files

Genome-wide copy number analysis and data quality control

DNA samples were sent for genome-wide copy number analysis using the Affymetrix OncoScan copy number array and run according to suggested manufacturer protocols. The OncoScan array is based on molecular inversion probe technology and provides comprehensive high-resolution copy number detection across the genome and at pan-cancer driver genes. OncoScan fluorescence array intensity (CEL) files were converted to OSCHP files using the hg19 reference (OncoScan_CNV.Ref103.na33.r1.REF_MODEL reference file included with the Affymetrix OncoScan Console software, version 1.3). Manual re-centering of samples was performed by adjusting the TuScan $\log_2 R$ using the OncoScan Console. Clonality analysis was performed with the Affymetrix OncoClone Composition tool.

Segmentation of NCI-MD and TCGA intensity files

For these samples, the Chromosome Suite Analysis (CHAS) was used for segmentation of intensity files at the default hyperparameters for output of segments with their copy number,

log₂R, and B-Allele frequency (BAF) information. For TCGA samples, Level 3 segmented files were retrieved from the firehose pipeline where a consistent version of reference (hg19) was used.

Quantifying GI, HRD, and Chromothripsis

Quantification of GI

Taking the output of segmentation results from the above for every sample in NCI-MD where we have copy number information for each segment, GI was defined by the ratio of the total length of regions with copy number other than two to a constant of 3.3E9, based on previous studies^{47,48}. We repeated this calculation for TCGA samples where we only selected cancer types with at least five AA samples.

Quantification of HRD

We identified five independent signatures to define somatic-level HRD (somatic HRD) across tumor samples, where four use copy number profiles and one uses the mutation profile of the tumor. We also used one signature to identify germline-level HRD (germline HRD) using germline variants in blood samples of the patients (detailed methods below). Here, we have described each one of them in detail.

I. Somatic HRD quantification

Based on loss of heterozygosity (LOH) regions Using the output of allele-specific segmentation, we identified and calculated a total sum of the number of LOH events, segments with only one allele, in each sample. Then, we normalized the value to be in the range [0,1] and termed it as LOH-HRD^{13,14}.

Based on telomere allelic imbalance (AIL) regions Again using the output of segmentation, we identified and counted the sum of regions with allelic imbalance, an unequal allele copy number, and extension to a sub-telomere without crossing the centromere. Again, we normalized the sum to be in the range [0,1] and termed the normalized sum as AIL-HRD.

Based on large-scale state transitions (LST) regions Here also, using the output of allele-specific segmentation, we identified and counted the total number of breakpoints between regions longer than 10 Mb after filtering out regions shorter than 3 Mb¹³. Again, we normalized the breakpoint counts to be in [0,1] and termed it as LST-HRD.

We defined the fourth method as $(\text{LOH-HRD} + \text{AIL-HRD} + \text{LST-HRD})/3$, scaled to 0-1, for each sample. The division by 3 puts the value in the range [0,1]. These four signatures were quantified and used in both NCI-MD and TCGA samples.

Exposures for each sample, the proportion of mutations assigned to mutation signature 3, known to be associated with HRD, was mined from mSignatureDB⁴⁹, a database of mutation signatures for more than 15,000 tumor samples from more than 73 projects, where only TCGA samples are considered for calculations.

II. Germline HRD quantification

Using the predicted pathogenic germline variants information in patients from TCGA ³⁰, we calculated the total number of pathogenic variants in HRD-genes in each sample (Supp Table 14) and performed a Fisher's exact test to identify whether AAs in comparison to EAs have a significantly higher frequency of pathogenic variants. We repeated this analysis for each HRD gene as well.

Purity and ploidy calculation

Using the OncoClone tool provided by Affymetrix, which uses the algorithm ASCAT ⁵⁰, we computed the purity and ploidy of samples from the NCI-MD cohort (Supp Table 2). Further, intratumor heterogeneity (ITH) was calculated using TuScan algorithm, a further extension of OncoClone.

Accessing variant calls of TCGA patients' blood samples from dbGAP

TCGA collection includes non-tumor biospecimen (blood samples were preferred if available, or adjacent non-tumor) for 10,224 patients with informed consent under the authorization of local institutional review boards of the sequence where whole-exome sequencing was performed ³⁰. We requested permission for these data from the database of Genotypes and Phenotypes (dbGaP) and after receiving permission, downloaded the variants from the controlled access part of the TCGA portal.

Quantification of chromothripsis

With an aim to identify whether an autosomal chromosome had undergone chromothripsis using SNCA profile data, we used four copy number based-hallmark traits of regions that underwent chromothripsis. Some of these hallmarks of chromothripsis have undergone an evolution since the first description, hence we used two partially overlapping hallmarks to identify chromothripsis based on the conventional ²⁰ and an alternative more recent ^{51, 52} description. Chromosomes that had all four hallmark properties were considered to have undergone chromothripsis.

We modeled the four hallmarks of chromothripsis via two tests for each sample. First, we filtered for chromosomes with significantly more events than the sample's background, derived from all other autosomes. Specifically, a chromosome must have a higher number of copy number events than the median number of copy number events per chromosome in the sample. Second, for every chromosome that passed the first test, the distance between the event breakpoints on the chromosomes should be significantly lower than the background distribution of copy number event breakpoints within the rest of the chromosomes. To this end, we tested whether the distances between the breakpoints of events of a given chromosome were lower than the background distribution of distances between the breakpoints of events on the rest of the chromosomes. If not, we removed the terminal event with a higher breakpoint distance from the penultimate and repeated the above step.

The above iteration was repeated for a chromosome until we found a region with greater than five events with significantly lower breakpoint distance (clustered, FDR-corrected $P < 0.1$), and the region comprised only one type of copy number event (oscillatory copy number state). We repeated the above steps with a single modification to model and detect CHTP based on the

recent definition, wherein a CHTP region can have two oscillatory copy number states or two types of copy number events.

Association of copy number change with expression

For this study, total RNA sequencing was performed for 56 out of 222 samples with SNCA profiles (31 LUAD & 25 LUSC). The association of copy number with expression was calculated via a one-tailed Wilcoxon rank-sum test, where samples were divided into two categories by thresholding on the median gene copy number to test, in a genome-wide fashion for each gene, whether samples with copy number higher than the gene median copy number in the cohort has expression significantly higher than the rest of the samples.

Focal and arm events by GISTIC

Generating a copy number map with focal-, arm- level events via GISTIC

The GISTIC algorithm was used to find recurrent regions of amplification, deletion, or LOH from the segmented file generated from CHAS. We used the following hyperparameter configuration throughout the study to find recurrent regions “*-genegistic 1 -smallmem 1 -broad 1 -brlen 0.5 -conf 0.90 -armpeel 1 -savegene 1*”. Based on this configuration, a gene GISTIC algorithm was used where *arm* level events are defined as aberrant regions with at least the length of half an arm, and regions below this threshold are defined as *focal*. The confidence level used to calculate the region was 0.90 and the q-value was the default of 0.25.

Unsupervised ancestry inference via principal component analysis (PCA) for NCI-MD cohort

Genotypes for 217,611 SNPs were generated from OncoScan OSCHP file via apt-tools for the samples from the NCI-MD cohort. We identified 46,217 SNVs variants likely to be ancestry-associated and not somatically acquired that are found to be present in at least 25% of the AAs or EAs in our cohort. In this matrix, where each row represents a patient and each column represents a SNP, we performed a PCA with rank two, constraining the number of principal components (PC) to two (Extended Data Figure 1). Next, we performed a classification using the two PCs using support vector classification (SVC) with a linear kernel to identify two classes. The predominant self-reported race in the class is assigned to be its identity. These two classes were then tested for concordance with self-reported ancestry.

Unsupervised ancestry inference via principal component analysis (PCA) for TCGA cohort

Genotype information of 906,601 SNPs from the SNP6 array performed on matched PBMC samples of TCGA patients called using BirdSeed, a SNP genotyping algorithm, were downloaded. We requested permission for these data from dbGaP and, after receiving permission, downloaded the variants from the controlled access part of the TCGA portal. To infer ancestry, methods similar to NCI-MD were employed, where after removing low-variance SNPs, we inferred 300,000 SNPs likely to be ancestry-associated that are found to be present in at least 25% of the AAs or EAs in our cohort. Following the methods described above for NCI-MD, we identified two classes of ancestry.

Statistical power analysis of TCGA samples from various populations

We observed a negative correlation between the FDR-corrected significance for AAs having higher GI and the proportion of AA samples included per cancer type, which was higher than expected when permuted a million times. (Spearman $Rho = -0.34$, $P < 0.15$; empirical $P < 1E-04$), suggesting that under-representation of samples from AAs is a limiting factor in terms of statistical power when comparing these two populations in certain tumor types in TCGA.

SNCA-gain and -loss based genomic instability (GI) analysis

For TCGA Pan-Cancer

We calculated SNCA-gain and SNCA-loss based GI and consistently observed both GI measures to be higher in AAs (Wilcoxon rank-sum $P < 5.2E-06$ and $P < 1.5E-06$). Further, the trend of higher GI was observed in 16 out of 23 cancer types for both SNCA-gain based and SNCA-loss based GI (Extended Figure 2A-B).

For NCI-MD LUSC

SNCA-gain and SNCA-loss based GI is calculated for LUSC from the NCI-MD cohort. We observed only SNCA-loss (Wilcoxon $P < 4.5E-06$) and not SNCA-gain (Wilcoxon $P < 0.34$) to be significantly higher in AAs ($P < 4.5E-06$ and < 0.34 , respectively).

Qualitative characterization of NCI-MD cohort tumor samples

Purity—the percentage of the tumor cell fraction within a sample—was successfully resolved in 194 out of 222 samples (Supp Table 2) where the mean purity was 34%. LUSC tumor samples

(38.5% mean purity) had a significantly higher (Wilcoxon $P < 0.009$) purity than LUAD (30.5%), consistent with the purity differences observed TCGA. The overall mean ploidy was 2.22.

Arm-level aberration frequency negatively correlated with the number of genes present on the chromosome arm (NCI-MD)

Broad level events across chromosome arms were quantified and plotted against the number of proteins expressing genes. We observed a general trend of negative correlation between the frequency of an aberration on a chromosome arm and the number of genes present on the same arm (median Spearman $Rho = 0.41$).

Data availability

Human TCGA cohort mutation data were derived from the publicly available mSignatureDB database: [<http://tardis.cgu.edu.tw/msignaturedb/>]. For the corresponding samples, copy number profiles, Level 3 segmented files were retrieved from the firehose pipeline [<https://gdac.broadinstitute.org/>] where a consistent version of reference (hg19) was used. The NCI-MD data were derived from patients enrolled in the ongoing NCI-MD Case-Control Study and all relevant data in this work is available on reasonable request, except for the TCGA pathogenic variant calls that required dbGaP controlled access and any sequence information that would make it possible to identify study subjects. Anonymized Level 3 segmented files for each sample, in addition of the raw files for copy number profiles of the NCI-MD patients and their corresponding expression profile, are deposited in dbGAP.

Code availability

We used open source R v3.6 throughout our work to generate figures. Wherever required, commercially available Adobe illustrator 23.0.3 (2019) was used to create figure grids. All the scripts for analysis and reproducing figures and panels are built in-house and are provided on github here: https://github.com/sanjusinha7/Scripts_MolCharAAvsEA.

Methods-only References

46. Enewold, L. et al. Serum concentrations of cytokines and lung cancer survival in African Americans and Caucasians. *Cancer Epidemiol Biomarkers Prev* 18, 215-222, doi:10.1158/1055-9965.EPI-08-0705 (2009).
47. Jamal-Hanjani, M. et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med* 376, 2109-2121, doi:10.1056/NEJMoa1616288 (2017).
48. Andor, N. et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* 22, 105-113, doi:10.1038/nm.3984 (2016).
49. Huang, P. J. et al. mSignatureDB: a database for deciphering mutational signatures in human cancers. *Nucleic Acids Res* 46, D964-D970, doi:10.1093/nar/gkx1133 (2018).
50. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* 107, 16910-16915, doi:10.1073/pnas.1009843107 (2010).
51. Yi, K. & Ju, Y. S. Patterns and mechanisms of structural variations in human cancer. *Exp Mol Med* 50, 98, doi:10.1038/s12276-018-0112-3 (2018).
52. Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J. & de Lange, T. Chromothripsis and Kataegis Induced by Telomere Crisis. *Cell* 163, 1641-1654, doi:10.1016/j.cell.2015.11.054 (2015).

Chapter 2: A systematic genome-wide mapping of oncogenic mutation-selection during CRISPR-Cas9 genome editing

Abstract

Recent studies have reported that genome editing by CRISPR–Cas9 induces a DNA damage response mediated by *p53* in primary cells hampering their growth. This could lead to a selection of cells with pre-existing *p53* mutations. In this study, employing an integrated computational and experimental framework, we systematically investigated the possibility of selection of additional cancer driver mutations during CRISPR-Cas9 gene editing. We first confirm the previous findings of the selection for pre-existing *p53* mutations by CRISPR-Cas9. We next demonstrate that similar to *p53*, wildtype *KRAS* may also hamper the growth of Cas9-edited cells, potentially conferring a selective advantage to pre-existing *KRAS*-mutant cells. These selective effects are widespread, extending across cell-types and methods of CRISPR-Cas9 delivery and the strength of selection depends on the sgRNA sequence and the gene being edited. The selection for pre-existing *p53* or *KRAS* mutations may confound CRISPR-Cas9 screens in cancer cells and more importantly, calls for monitoring patients undergoing CRISPR-Cas9-based editing for clinical therapeutics for pre-existing *p53* and *KRAS* mutations.

Introduction

CRISPR enables targeted gene-disruption and editing, a powerful technology that expands our understanding of fundamental biological processes¹. Beyond its impact on biological research, CRISPR-based approaches have been considered for various applications in medicine, from reparative editing of primary cells to the development of new strategies to treat a variety of genetic diseases, including cancer. However, several clinical trials based on CRISPR technology

have been deferred due to significant potential risks, including off-target effects^{2,3,4}, generation of unexpected chromosomal alterations⁵ and potential immunogenicity⁶. Other studies have demonstrated that double stranded breaks (DSBs) induced during CRISPR-Cas9-based gene knockout (CRISPR-KO) can lead to DNA damage response, whose level is associated with the copy number of the targeted gene⁷⁻¹⁰.

Recent studies have shown that the DNA damage response following CRISPR-KO can be mediated by *p53*, a tumor-suppressor gene mutated in over 50% of all human cancers^{11,12}. Genome-wide CRISPR screening in immortalized human retinal pigment epithelial (RPE1) cells¹² revealed that a *p53*-mediated DNA damage response, followed by cell cycle arrest, is induced upon generation of DSBs by the Cas9 endonuclease, favoring the survival of cells that have inactivated the *p53* pathway. Most recently, a study showed that exogenous expression of Cas9 can also activate this *p53*-mediated DNA damage response¹³. While these studies indicate that CRISPR-Cas9 genome editing techniques may select for *p53* mutated cells^{11,12,13}, several outstanding questions remain unaddressed: First, since most of these *p53* studies have involved only a small number of primary or stem cells^{11,12}, it is unclear whether *p53* selection can happen broadly across multiple different cell types including transformed cancer cells. Second, it is not clear whether stronger *p53* selection can happen when certain genes or parts of the genome are targeted, or the level of selection is more homogenous regardless of the genes being edited. And finally, it remains to be investigated whether this selection is limited to *p53* only or that other cancer driver genes can also be selected for during CRISPR-Cas9 genome editing.

To address these questions, here we employ a computational framework coupled with experimental validations to conduct a comprehensive evaluation of each cancer driver mutation selection associated with CRISPR-Cas9. We first demonstrate that CRISPR-KO-induced mutant

p53 selection can be observed in transformed and non-transformed cells of diverse lineages via both lentivirus and ribonucleoprotein-based Cas9 delivery. More importantly, we systematically characterized mutation selection in other cancer driver genes during CRISPR-Cas9 identifying that *KRAS* mutants can also be selected for, as demonstrated in large-scale genetic screens and Cas9-expressing cell lines. We further identified the underlying pathways that are likely to mediate this selection.

Results

CRISPR-Cas9 gene-knockouts selects for p53 mutations in a vast variety of transformed and non-transformed cell types

We first sought to address two important gaps in our understanding of CRISPR-KO-driven mutant *p53* selection – firstly, we wanted to investigate whether this selection generalizes across cell types. Secondly, we wanted to understand what type of sgRNAs, genes and gene-networks drive this selection. We analyzed the *DepMap*¹⁵ genome-wide gene essentiality data across 248 cancer cell lines (**Table S1**), where both CRISPR-Cas9 (*AVANA*¹⁰) and shRNA-based (*Achilles*¹⁵) genetic screens were conducted. We searched for genes whose CRISPR-Cas9-based knockout (CRISPR-KO) reduced cell viability more (i.e. more essential) in *p53*-wildtype (WT; N=75) than *p53*-mutant (N=173) cell lines, but do not exhibit such differential essentiality in the shRNA-based screens (Methods). The KO of such genes may select for *p53* mutants specifically during CRISPR-Cas9 editing. In the CRISPR-Cas9 screen, we find many more genes (981) that are more essential in *p53*-WT vs *p53*-mutant cell lines, compared to the genes that are more essential in *p53*-mutant cells (237 genes). In contrast, the numbers of such differentially essential genes in the shRNA screens were balanced (~1500 each). Such significantly different patterns

between CRISPR-Cas9 and shRNA screens (**Figure 1a** left panel, Chi-squared test $P < 1.4E-284$) points to a bias that knockout/knockdown of a gene is more likely to impair the fitness of *p53*-WT cells specifically with CRISPR-Cas9 but not with shRNA. Potential confounding factors including gene copy number, functional impact of *p53* variants and phenotype difference between gene knockout/knockdown are discussed and controlled for in this analysis (**Supp. Note 1, Figure S1**).

Among the 981 genes that are more essential in *p53*-WT cells with CRISPR-KO, 861 genes (87%) do not exhibit this differential essentiality in shRNA screens. We hence termed these *CRISPR-specific differentially essential positive (CDE+)* genes (**Figure 1a** right panel; genes listed in **Table S2A**). We find that these CDE+ genes are preferentially located in chromosomal bands containing common fragile sites (CFSs; hypergeometric $P < 2.3E-4$, **Figure 1b, Table S3**), which are prone to replicative stress, fork collapse and DNA breaks that cause genomic instability¹⁶. As CRISPR-KO could induce kilobase-scale structural alterations near the targeted site¹⁷, this finding suggests that CRISPR-targeting near CFSs may enhance DNA damage, promote the *p53*-dependent cell death response and provide a selective advantage to *p53* mutant cells. The sgRNAs of the CDE+ genes also tend to target highly accessible chromatin (hypergeometric $P < 0.02$; Methods), thus inducing a strong damage response as recently reported¹⁸. The top pathways enriched within CDE+ genes include DNA damage response, DNA repair and Fanconi anemia (FA; hypergeometric test adjusted $P < 0.01$, **Table S2B**). This is consistent with the recent report that the FA pathway is involved in repairing Cas9-induced DNA double-strand breaks (DSBs)¹⁹ and that their KO may further enhance DNA damage.

Analogous to CDE+, we defined CDE- genes, which are more essential in *p53*-mutant (vs WT) cells with CRISPR-KO, but not showing such difference in shRNA screens (185 genes,

right panel of **Figure 1a**). CDE- genes are involved in cellular processes that engage *p53*, including mitotic checkpoints, DNA replication and cell cycle (**Table S2B**, **Figure 1d**, hypergeometric test adjusted $P < 0.1$), with the top hit being the key cell cycle regulator *CDKN1A*^{11,12} (a.k.a. *p21*, Wilcoxon rank-sum $P < 1.85 \times 10^{-8}$, **Figure 1c**). Transiently inhibiting CDE- genes during CRISPR-KO may mitigate *p53* mutation selection and could be of interest from a translational point of view. Top CDE+/- genes are highlighted in **Figure 1c**. We repeated this CDE+/- identification process using an independent CRISPR-Cas9 screen in 326 cancer cell lines²⁰ and observed concordant results (**Supp. Note 2**, **Figure S2**).

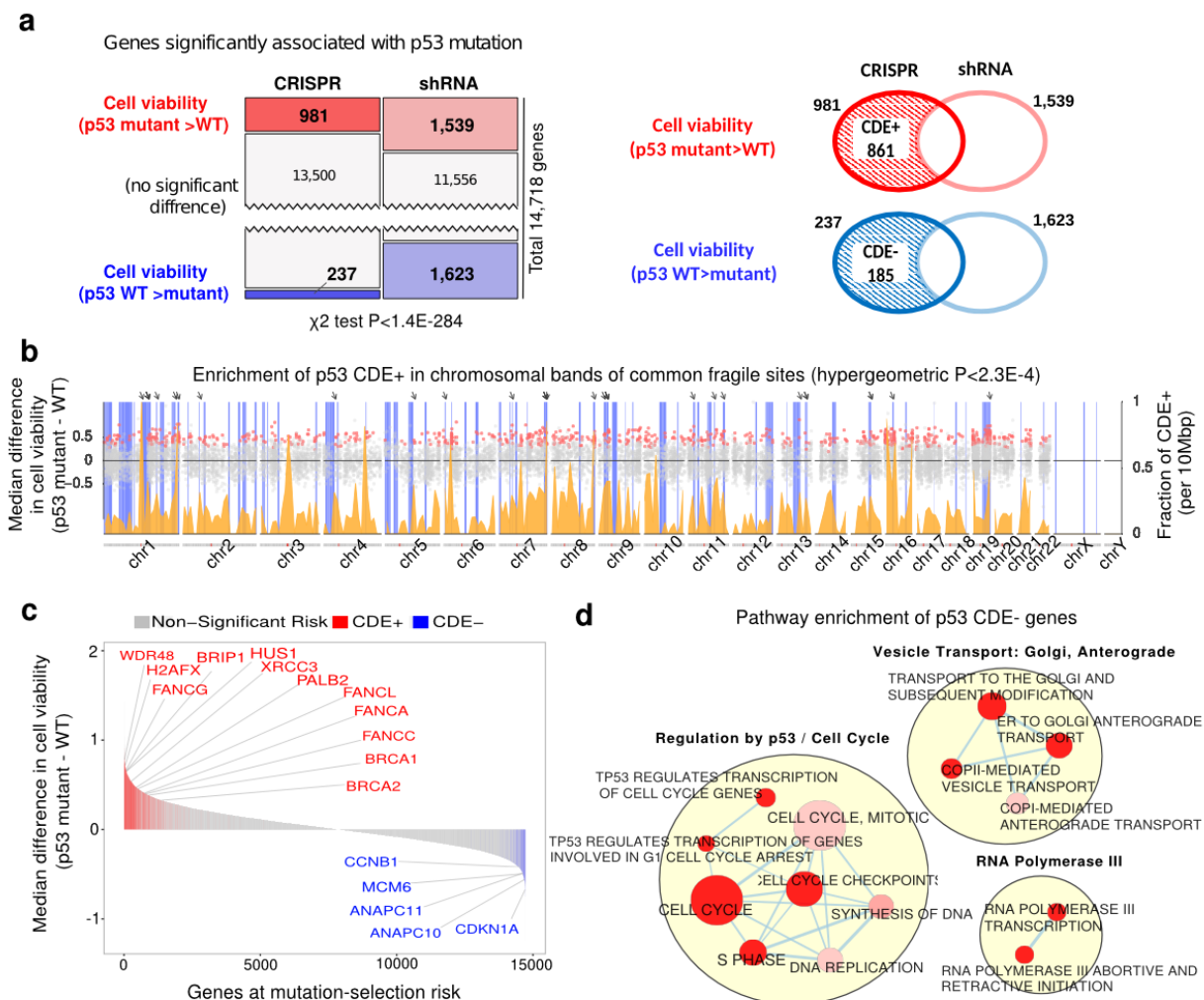


Figure 1. A genome wide view of *p53*-mutant selection. (a) Upper panel: number of genes whose essentiality is significantly associated with *p53* mutation status in CRISPR and shRNA screens (one sided Wilcoxon rank-sum has been performed with FDR threshold of 0.1). Lower panel: the definition of CDE+ and CDE- genes. (b) Enrichment of *p53* CDE+ genes in common fragile sites (CFSs). The x-axis denotes the chromosomal position; the scatter plot (y-axis on the left-hand side) shows the difference of median post-CRISPR-KO cell viability values in *p53* mutant vs *p53* WT cell lines for *p53* CDE+ genes (red dots) and all other genes (grey dots); the density plot (colored orange, y-axis on the right-hand side) shows the fraction of *p53* CDE+ genes among all genes per DNA segments of 10 Mbp along the genome; the vertical blue bars indicate the chromosomal bands of CFSs, and prominent sites where peaks of high CDE+ gene density coincide with CFSs are marked by arrows on the top. (c) The distribution of predicted level of CRISPR-Cas9 *p53*-mutant selection across the genome. Significant CDE+ genes that are part of FA pathway are marked in red and significant CDE- genes that are part of cell cycle regulation in blue. (d) Visualization of the pathways enriched for *p53* CDE- genes where significance is calculated using the GSEA method as implemented in the R package fgsea [21]. Only significantly enriched pathways (FDR<0.1) specific to CRISPR (and not in genes showing differential essentiality in the shRNA screens are shown). Pathways are depicted as nodes whose sizes correlate with pathway lengths and colors represent enrichment significance (the darker, the more significant). Pathway nodes are connected and clustered based on their functional similarities, and higher-level functional terms are given for each of the clusters (Methods). For clarity, only the largest clusters are shown.

We next performed our own CRISPR-Cas9 essentiality screen, employing CRISPRi-based essentiality screens as a control in a pair of *p53*-isogenic MOLM13 leukemia cell lines

(WT and *p53* R248Q mutant). We used a deep (10 guides per gene) and focused sgRNA library targeting top *p53* CDE+ and CDE- genes (Methods; **Figure 2a**, details in **Supp. Note 10, Table S4**). Here, we observed in the CRISPR-Cas9 screen that the CDE+ genes are more essential in *p53*-WT vs mutant cell, and *vice versa* for the CDE- genes (Wilcoxon signed-rank test $P < 0.08$ and $P = 0.03$ for CDE+/- genes respectively). Reassuringly, we do not see such differential essentiality in the CRISPRi screens (Wilcoxon signed-rank $P = 0.32$ and 0.29 ; Top 10% CDE+/- genes are depicted in **Figure 2b**).

To further assess whether such selection effects can be observed in non-transformed cells, we next tested and observed that indeed our *p53* CDE+ genes have a higher essentiality in WT vs isogenic-mutant cells in published CRISPR-Cas9 [12] but not shRNA [47] genome-wide screens performed in non-transformed RPE1 cells (**Figure S3**, screens quality control discussed in **Supp. Note 9** and **Figure S4**). This finding is further confirmed by mining seven CRISPR-KO genome-wide screens²², including two *p53*-null and five *p53*-WT RPE1 cells screens (**Figure S5**, **Supp. Note 3**).

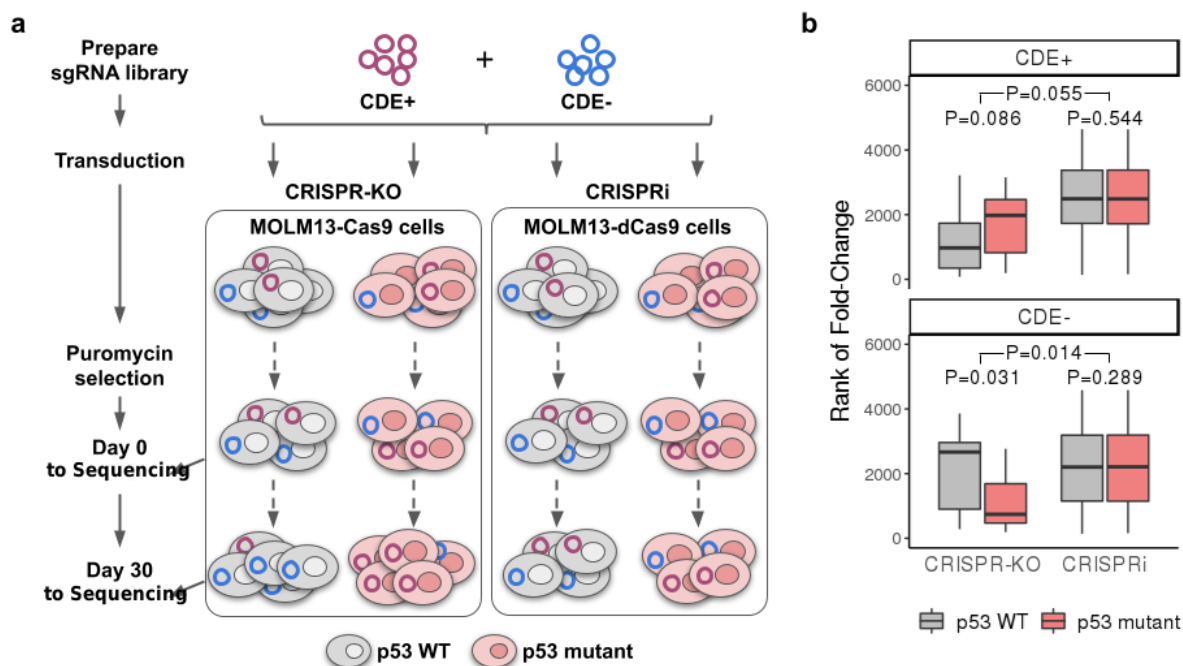


Figure 2. Validation of *p53* CDE genes in isogenic MOLM13 cell lines via pooled CRISPR screens. **(a)** A flowchart showing the experimental procedure of CRISPR-KO and CRISPRi screening of pooled *p53* CDE \pm genes in a pair of *p53*-isogenic MOLM13 cell lines. See Methods for details. **(b)** The day 30 to day 0 fold-change (converted to rank) of reads corresponding to the sgRNAs for *p53* CDE $+$ genes (upper panel) and CDE $-$ genes (lower panel), in *p53* WT MOLM13 cells (gray boxes) vs the isogenic *p53* mutant cells (red boxes) for the CRISPR-KO and CRISPRi screenings, respectively. The bottom P values are for Wilcoxon signed-rank tests comparing *p53* WT and mutant cells, the upper ones are P values of non-parametric tests comparing the difference of *p53* mutant and WT rank values between CRISPR-KO and CRISPRi experiments. In the boxplots, the center line, box edges and whiskers denotes the median, interquartile range and the rest of the distribution in respective order, except for points that were determined to be outliers using a method that is a function of the interquartile range, as in standard box plots.

A competition assay shows selection for p53 mutant over wildtype cells following CRISPR-Cas9 knockout of CDE+ genes

To test whether the CRISPR-KO of CDE+ genes leads to selection of *p53* mutant cells in a competitive setting²³⁻²⁵, we silenced the top five predicted CDE+ genes using CRISPR-Cas9 and CRISPRi in the *p53*-isogenic MOLM13 cells (Methods). Following a lentiviral sgRNA transduction, the WT and mutant cells were mixed at an initial ratio of 95:5, and monitored by flow cytometry for up to 25 days (illustrated in **Figure 3a**; **Table S5A**). Silencing 2/5 CDE+ genes (*NDUFB6* and *NDUFB10*) induced a strong progressive *p53* mutant enrichment of up to five folds over WT at day 25 specifically in CRISPR-KO, across several independent sgRNAs and not for NTC (**Figure 3b**, blue lines). No inverse enrichment in *p53* WT cells was observed in the competitive assays involving the three other CDE+ genes (**Figure S6**). We observed that sgRNAs targeting *NDUFB6* induced significantly higher DNA damage compared to NTC-treated cells specifically in *p53* WT cells (**Figure S7**, despite editing efficiency being higher in the mutant cells as shown in **Figure 4c**), demonstrating that the DNA damage was not just due to Cas9 expression. This may partly explain their selective competitive advantage upon the CDE+ gene KO. Testing the robustness of this competitive selection advantage for *p53* mutant cells, we repeated the CRISPR-KO competitive assay for a larger number of 18 top CDE+ genes with up to 4 unique sgRNAs per gene and monitored the assay up to 15 days (**Table S5B**). Using the non-targeting sgRNA as a baseline, we observed the competitive outgrowth of *p53* mutant cells for 15 out of 28 sgRNAs and 10 out of 18 CDE+ genes tested (**Figure 3c**).

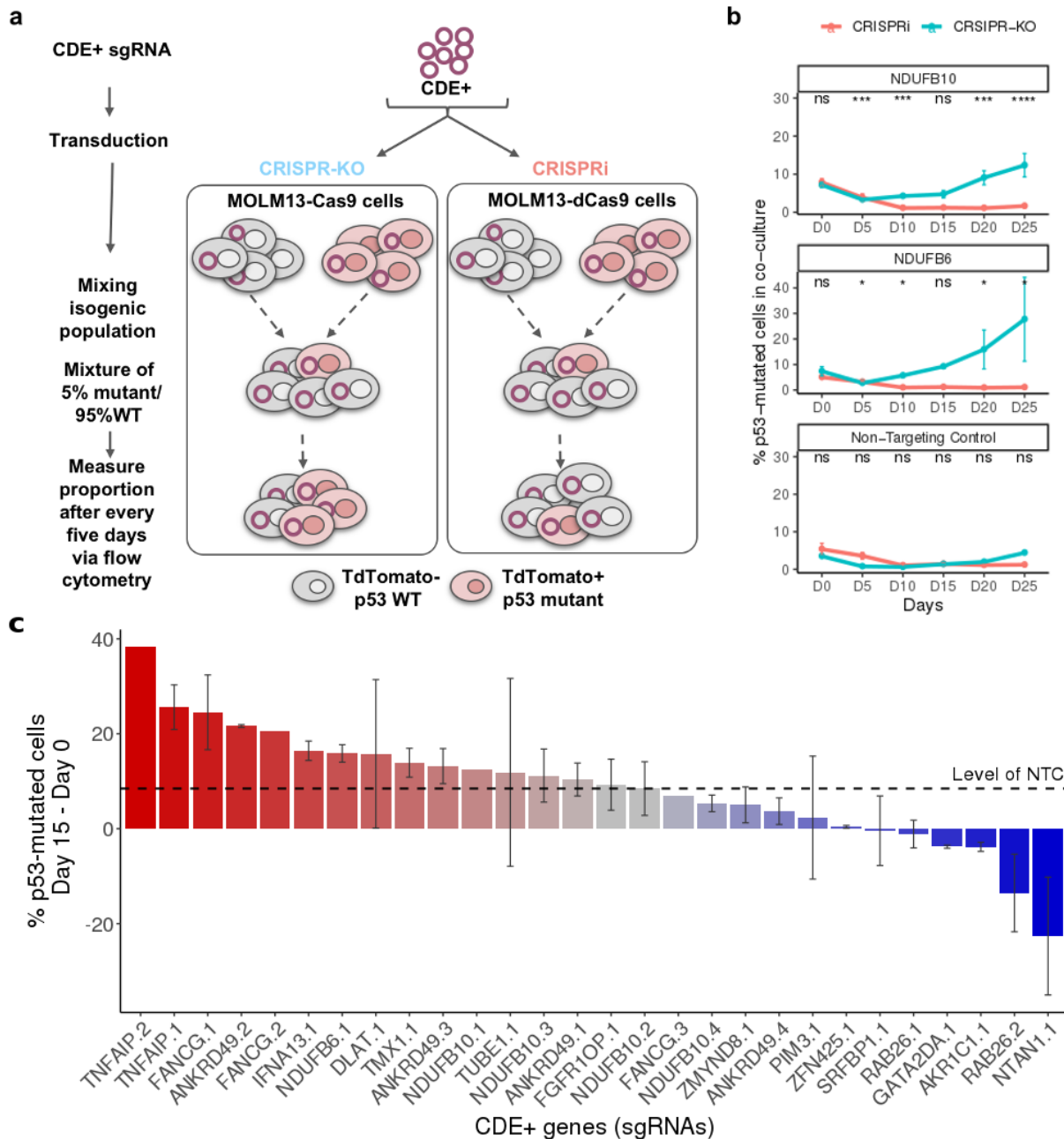


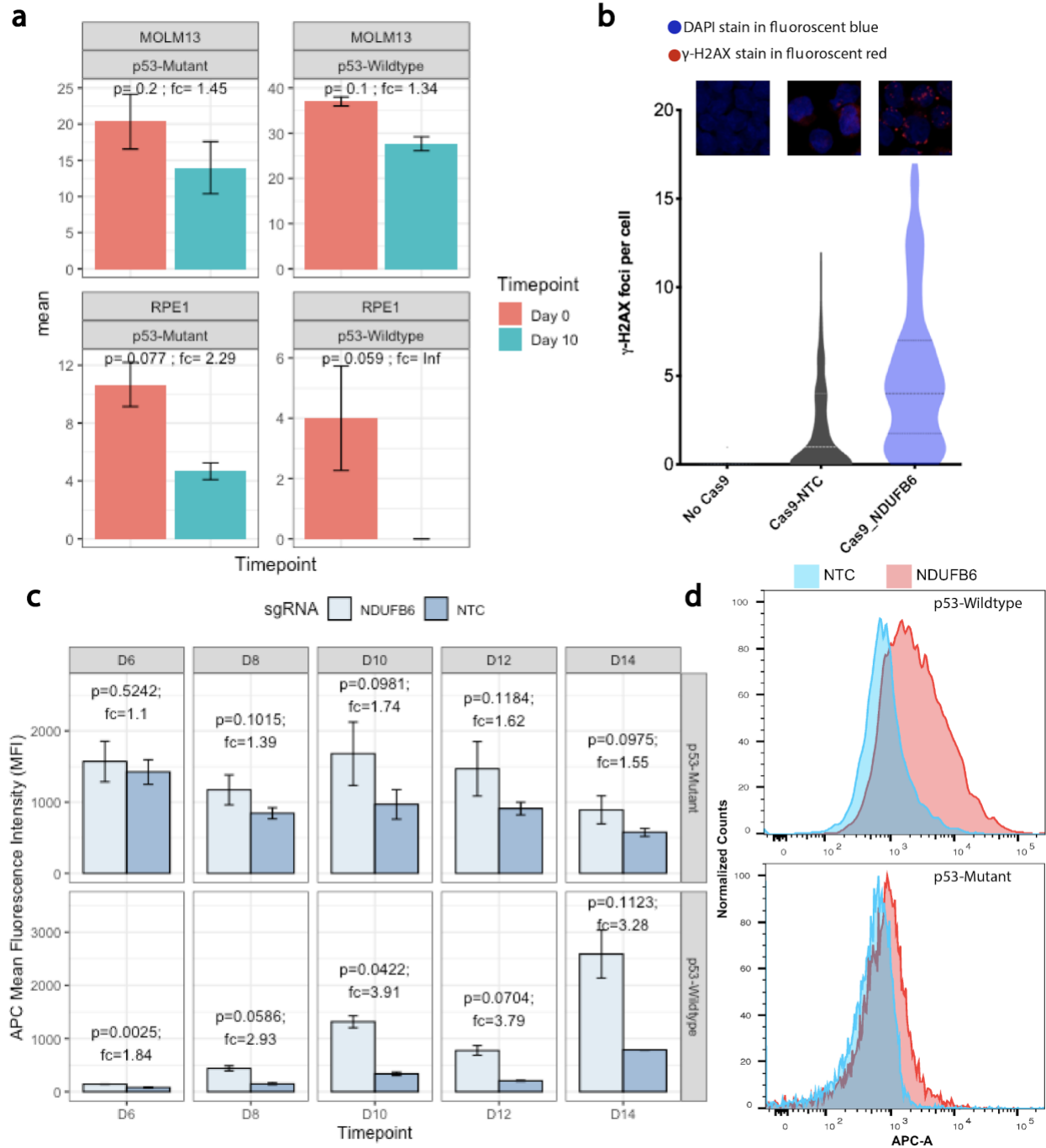
Figure 3. Selection for *p53* mutant cells under CRISPR-Cas9 knockout of CDE+ genes in a co-culture of *p53* WT/mutant cells. (a) An illustration showing the experimental design of the competition assay where isogenic *p53* WT/mutant MOLM13 cell lines were mixed with a ratio of 5:95 and top *p53* CDE+ genes were knocked out by CRISPR-Cas9. Population ratio was monitored for 25 days at a five-day interval starting from the day of sgRNA transduction. **(b)**

Change in the percentage of *p53* mutant cells in the *p53* mutant-WT cells (Y-axis) co-culture with time (X-axis, number of days in co-culture), under the CRISPR-KO or CRISPRi of individual selected top *p53* CDE+ genes or with non-targeting control sgRNA. The p-values are calculated using two-sided Wilcoxon Rank Sum tests. (c) The difference of the percentage of *p53* mutant cells between Day 15 and Day 0 in co-culture (Y-axis), under the CRISPR-KO of a larger set of top *p53* CDE+ genes (X-axis, by individual sgRNAs, specified by number suffixes after gene symbols). Error bars represent standard error across replicates for each sgRNA. The horizontal dashed line represents the value for non-targeting control sgRNA (NTC).

p53 mutation selection phenomena extends to transient knockout and primary cells

We next asked whether our top CDE+ genes may also select for *p53* mutants under CRISPR-Cas9 transient knockout. We delivered Cas9 and the sgRNA as a ribonucleoprotein (RNP). We observed that upon Cas9-RNP mediated transfection of a sgRNA targeting our top CDE+ gene from our pooled and competition assays, *NDUFB6* (see Methods), there was a higher loss of edited cells in the *p53* WT vs isogenic *p53* mutant MOLM13 cells over 10 days of culture, as measured by change in ICE scores (**Figure 4a top panel**). Using an orthogonal method of proliferation monitoring by dye-dilution²⁶, we observed that there was a progressive slowing down in cell proliferation of *p53* WT, but not *p53* mutant MOLM13 cells upon Cas9-RNP based KO of *NDUFB6* vs respective non-targeting controls (**Figure 4b,c**). Similar to the lentiviral system, this is likely due to the DNA damage induced by the *NDUFB6* sgRNA compared to the Cas9 only or Cas9 with NTC controls in *p53* WT cells (**Figure 4b** and **S8**). We repeated this transient knockout in non-transformed cells (RPE1) and consistently observed an increased loss of edited *p53* WT over *p53* mutant cells (**Figure 4a, bottom panel**). Notably, we

also observed a selection of *p53* mutant over WT in patient tumors profiles (TCGA) based on the copy number alteration patterns of CDE+ genes (details in **Supp. Note 11A**).



determined in *p53* mutant and wildtype isogenic pair of MOLM13 (top panels, transformed cells) and RPE1 cells (bottom panels, primary cells) using ICE protocol (see Methods) at day 0 (orange) and day 10 (green) after Cas9-RNP-sgRNA nucleofection. Differences in day 0 compared to day 10 editing efficiency can be used as a measure of relative fitness of edited compared to non-edited cells. The p-values are calculated using two-sided Wilcoxon Rank Sum tests. In the boxplots, the center line, box edges and whiskers denote the median, interquartile range and the rest of the distribution in respective order, except for points that were determined to be outliers using a method that is a function of the interquartile range, as in standard box plots.

(b) DNA damage is quantified from gamma H2AX staining images and measured by gH2AX staining in *p53* wildtype MOLM13 cells with no Cas9, Cas9 + sgRNA for a non-targeting control (NTC) or *NDUFB6*. gH2AX foci (y-axis) in all three conditions (x-axis) are enumerated in the violin plot. **(c)** Mean fluorescence intensity of the CellTrace™ dye (APC) in MOLM13 *p53* mutant (top panel) vs wildtype cells (bottom panel) is shown for NTC (light blue) or *NDUFB6* targeting sgRNAs (dark blue). CellTrace™ APC fluorescence is inversely correlated with proliferation. The error bars denote standard error (mean +/- standard deviation) across three replicates. The p-values are calculated using a two-sided t-test given a small number of data points (n=3). **(d)** Proliferative effects of *NDUFB6* editing in RNP-transfected *p53* mutants compared to wildtype cells. A histogram of MOLM13 *p53* wildtype (top panel) cells transfected with an NTC or a *NDUFB6* sgRNA are shown with the fluorescence intensity of the CellTrace™ dye (APC) on the x-axis. Similarly, MOLM13 *p53* mutant cells are plotted in the bottom panel.

KRAS mutant cell lines exhibit selection advantage in large-scale genetic screens

To determine whether additional cancer driver mutations may be selected for following CRISPR-KO, we focused on a list of 61 cancer driver genes from Vogelstein *et al.*²⁷ that are

mutated in at least 10 of the cell lines screened in the AVANA¹⁰ and Achilles¹⁵ datasets. For each of these cancer genes, we identified the differentially essential genes between its WT and mutant cell lines in the CRISPR-Cas9 (AVANA) and shRNA (Achilles) screens, as described above for *p53*. We ranked the cancer genes by the significance of skewness in the numbers of differentially essential genes from CRISPR-Cas9 vs shRNA screens similar to that shown in **Figure 1a** for *p53* (with Fisher's exact tests, Methods; results shown in **Figure 5a** and **Table S6**). The mutants of these genes may be selected for during CRISPR-KO, as their WT cells are overall more vulnerable during CRISPR-KO compared to the mutants. We term these genes "(potential) CRISPR-selected cancer drivers" (CCDs). The top significant CCD in addition to *p53* is the oncogene *KRAS*. Like for *p53*, potential confounding factors including copy number were controlled for (**Supp. Note 1, Figure S1b**), and there is no significant correlation between the mutation profiles of *KRAS* and *p53* (Fisher's test $P=0.67$), suggesting that *KRAS* might be a CCD independent of *p53*. We thus next focused on investigating the selection of mutant *KRAS* as another major CCD.

KRAS is a major oncogene whose gain of function mutation is known to activate various DNA repair pathways and may override the trigger of cell death upon DNA damage^{28,29}, supporting its role as a CCD. We computationally identified the CDE+ and CDE- genes of *KRAS* in a similar way described above for *p53* (**Figure 5b, Table S2A**). *KRAS* has high numbers of CDE+/- genes, while only very few *KRAS* mutation-associated genes are identified in the shRNA screen. The predicted median mutant selection levels are comparable to those of *p53* (**Supp. Note 4**), i.e. the CRISPR-KO of its CDE+ genes is likely to drive comparable levels of mutant selection as the KO of the CDE+ genes of *p53*. Fourteen genes are CDE+ genes of both *p53* and *KRAS*, and thus their CRISPR-KO may impose considerable selection for both *KRAS* and *p53*

mutants (**Table S2A**). Consistent with the knowledge of downstream pathways regulated by activated *KRAS*^{28,29}, its CDE- genes are significantly enriched for DNA DSB repair pathways (FDR<0.02, Methods, visualized in **Figure 5c**, **Table S2E**).

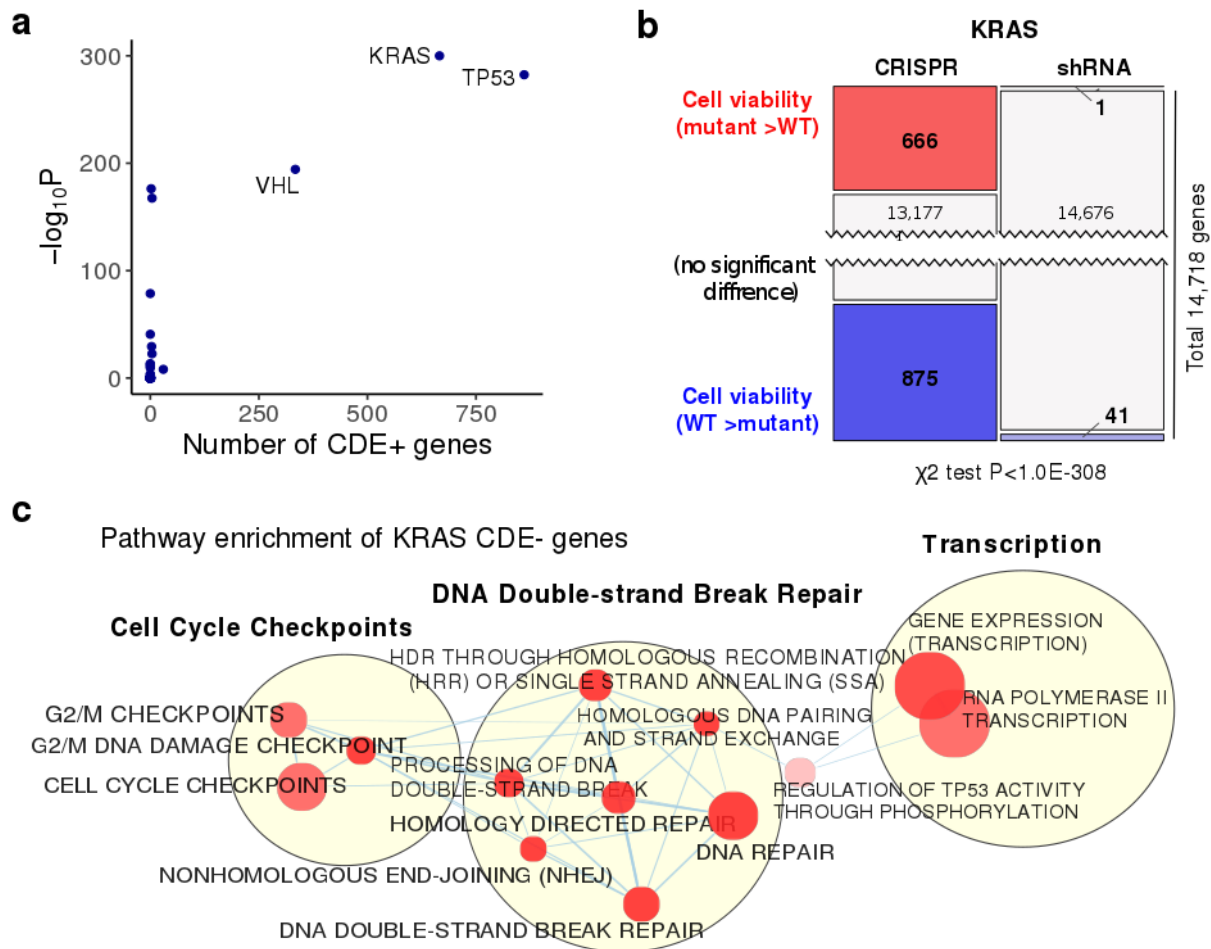


Figure 5. Large-scale genetic screening identifies *KRAS* as a second major cancer driver whose mutation can be potentially selected for by CRISPR-Cas9. (a) A scatter plot showing the number of identified CDE+ genes (X-axis) and the negative log10-transformed P values of Fisher's exact test (Y-axis) testing for the imbalance in the number of differentially essential

genes in CRISPR and shRNA screens for the 61 major cancer driver genes from Vogelstein *et al.*²⁷. *p53* and *KRAS* are identified as the top two significant cancer genes with the higher number of CDE+ genes. **(b)** The number of genes whose essentiality is significantly associated with *KRAS* mutational status in CRISPR and shRNA screens (one sided Wilcoxon rank-sum has been performed with FDR threshold of 0.1). **(c)** Visualization of pathways enriched for *KRAS* CDE- genes where significance is calculated using the GSEA method as implemented in the R package *fgsea*²¹. Only significant pathways (FDR<0.1) specific to CDE and not to the genes showing differential essentiality in the shRNA screens are included. Pathways are shown as nodes whose sizes correlate with pathway lengths and colors represent the significance of their enrichment (the darker the more significant). Pathway nodes are connected and clustered based on their functional similarities, and higher-level functional terms are given for each of the clusters (Methods). For clarity, only the largest clusters are shown.

Similar to *p53*, we next performed our own CRISPR-Cas9 and a control CRISPRi gene essentiality screens, but on a smaller scale, in a pair of isogenic *KRAS* WT and *KRAS* G12D mutant MOLM13 cell lines using a sgRNA library targeting top *KRAS* CDE+ and CDE- genes (details in **Supp. Note 10, Table S7**). Here, we observed that the *KRAS* CDE+ genes are more essential in WT than mutants and *vice versa* for CDE- genes specifically in CRISPR-Cas9 screens (Wilcoxon signed-rank test $P<0.074$ and $P<0.042$ for CDE+/- respectively, **Figure 6a left panel**), but not in CRISPRi screens (Wilcoxon signed-rank $P<0.22$ and $P<0.49$; **Figure 6a right panel**). Similar results were obtained from analyzing published genome-wide CRISPR-Cas9³⁰ and shRNA genetic screens³¹ performed in a different pair of *KRAS* isogenic cell lines (WT and G12D mutation in DLD1 cell line; **Figure 6b**). Similar to *p53*, we also observed a

selection of *KRAS* mutants in patient tumor profiles (TCGA³²) based on the copy number alteration patterns of its CDE+ genes (details in **Supp. Note 11B**).

A competition assay shows selection for *KRAS* mutant over wildtype cells following CRISPR-Cas9 knockout of *KRAS* CDE+ genes

To test whether, like the *p53* case, CRISPR-KO of *KRAS* CDE+ genes can confer a selective advantage to *KRAS* mutant over WT cells in co-culture, we conducted a similar competition assay using a pair of WT and *KRAS* G12D mutant isogenic MOLM13 cell lines. As in the experiment for *p53*, we mixed the WT and *KRAS* mutant cells at an initial ratio of 95:5 following *KRAS* CDE+ sgRNA transduction and monitored the population for 15 days to track the percentage of *KRAS* mutant cells (TdTomato+) with flow cytometry (Methods; **Table S8**). A total of 10 *KRAS* CDE+ genes were tested, in addition to a non-targeting control (NTC). In the control group, the *KRAS* mutant cell fraction decreased with time, indicating that the mutant cells have lower baseline fitness levels than the WT cells. In comparison, in 8 out of 10 CDE+ genes tested, there is a gain in fitness of the mutant cells (**Figure 6c**; Methods), testifying that even though the *KRAS* mutant cells have a lower baseline fitness level, the CRISPR-KO of the majority of CDE+ genes can enhance their fitness and in a subset of cases lead to selective outgrowth of *KRAS* mutant over WT cells in a mixed population.

Cas9 expression in cancer cell lines selects for *KRAS* mutations

Multiple studies have reported a higher editing efficiency of Cas9 in *p53* mutated versus *p53* WT cell lines^{11,12,14,13}. We first asked if this may also extend to *KRAS* as an equally important CCD. Analyzing induced exogenous Cas9 activity in 1601 cancer cell lines from DepMap (1375 and 226 *KRAS* WT and mutant, respectively)¹⁰, we find that, like *p53*, Cas9

activity is significantly higher in *KRAS* mutant cells than in *KRAS* WT cells ($P=2.9E-05$, **Figure 6d**; Methods). We repeated the above analysis modeling Cas9 activity vs *KRAS* status adjusting for *p53* status in a linear model, yielding concordant findings ($P=2.43E-04$). Importantly, across all the 61 cancer driver genes we analyzed above from Vogelstein et al.²⁷, *KRAS* and *p53* are the only ones showing such a significant difference in Cas9 activity between WT and mutant cells after FDR correction ($FDR=9.1E-04$ for *KRAS* & $7.1E-06$ for *p53*, **Figure 6e**). This further shows that in addition to *p53*, *KRAS* WT status can also hamper the efficiency of CRISPR-Cas9.

Based on the above findings and a recent report¹³ of selection for *p53* mutant due to DNA damage upon Cas9 expression (without sgRNA), we asked whether DNA damage induced by Cas9 alone can also lead to a mutation selection of *KRAS* and/or other cancer drivers. To this end, we re-analyzed deep sequencing profiles from 42 Cas9-expressed vs matched parental (i.e. without Cas9) cell lines (Methods) from Enache *et al.*¹³, and identified a total of 9 cases involving 5 unique *KRAS* mutations, occurring in 7 different cell lines with moderate to high Cas9 activity (Methods). Seven out of these 9 cases show increased mutant allele frequency after induced Cas9 expression (Wilcoxon signed rank test $P=0.027$, **Figure 6f**). Four of the 5 *KRAS* mutations are missense mutations; the other mutation is an intronic mutation occurring 100bp from the splicing site. This mutation is not present in the parental cell lines but emerges independently after Cas9 expression in four different cell lines. While the results suggest that CRISPR-Cas9 may select for *KRAS* mutations, the functional role of these mutations needs to be interpreted with caution. Among the 61 cancer driver genes from Vogelstein *et al.*²⁷, *KRAS* is a top gene (ranked the second) along with *p53* (ranked fourth) that show significant mutant sub-clonal expansion (**Figure 6g**). Notably, the top genes identified to be involved in mutant sub-

clonal expansions have a significant overlap with our previously identified top CCD genes (Fisher's exact test $P=0.04$).

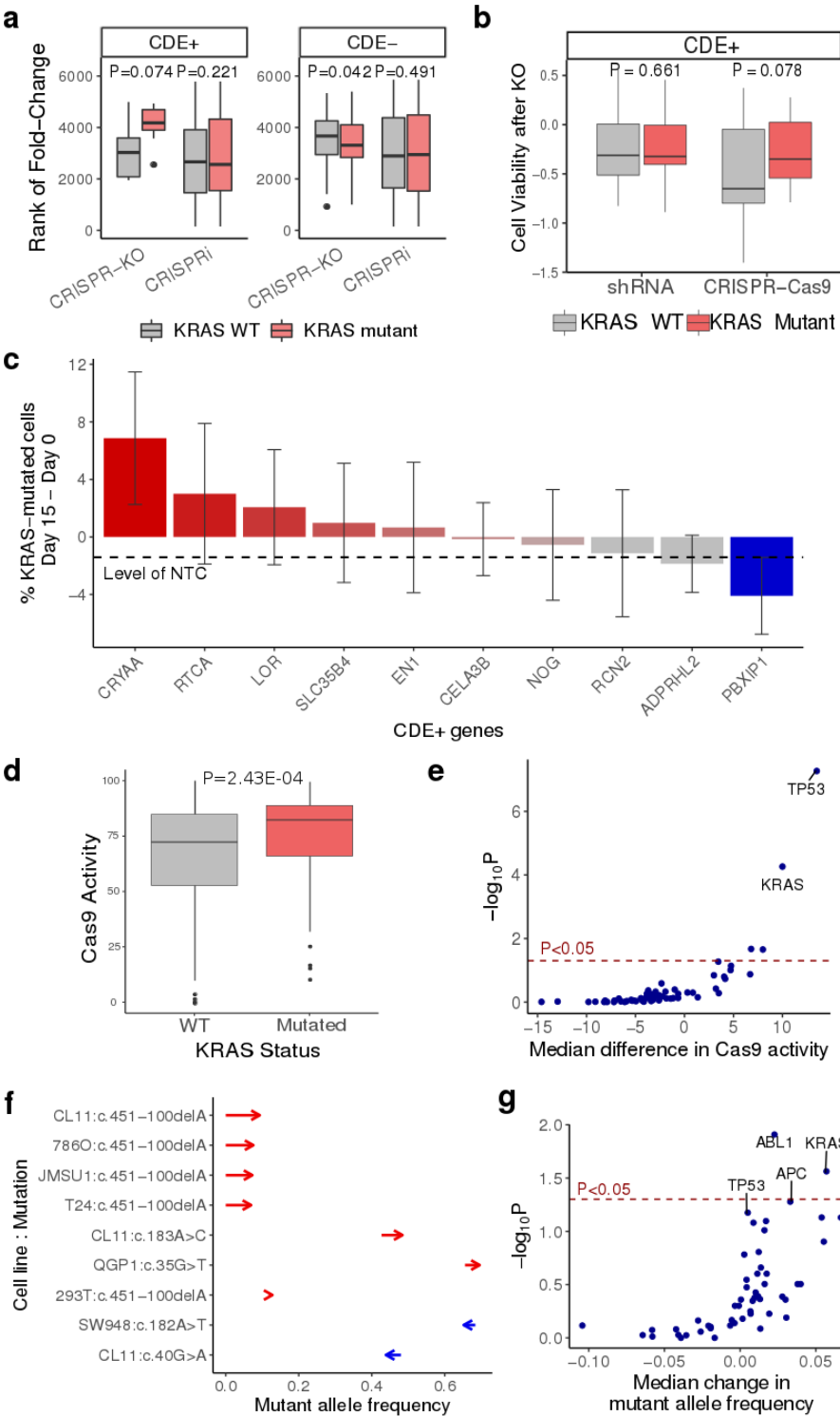


Figure 6. Beyond *p53*: experimental evidence identifying *KRAS* as another major cancer driver whose mutation can be potentially selected for by CRISPR-Cas9. (a) CRISPR-Cas9 and CRISPRi screens of the top *KRAS* CDE gene knockouts were performed in isogenic MOLM13 and MOLM13-*KRAS*-G12D cell lines. The box plot shows the trend that the sgRNAs of the *KRAS* CDE+ genes are more depleted in *KRAS* WT cells vs *KRAS* mutant cells and vice versa for *KRAS* CDE- genes in CRISPR-Cas9 screens, but there is no such trend in the CRISPRi screens. The P values shown are of one-tailed Wilcoxon signed-rank tests. (b) Analysis of published genome-wide CRISPR-Cas9 and shRNA screens in *KRAS*-isogenic DLD1 cell line. The box plot shows the trend that the CRISPR-KO of *KRAS* CDE+ genes reduces cell viability more in *KRAS* WT cells than *KRAS* mutant cells, while there is no such trend in the shRNA screen. The P values of one-tailed Wilcoxon signed-rank tests are shown. (c) The difference in the percentage of *KRAS* mutant cells between Day 15 and Day 0 in co-culture (Y-axis), under the CRISPR-KO of different *KRAS* CDE+ genes (X-axis). Error bars represent standard error. NTC: non-targeting control sgRNA. (d) A box plot showing the comparison of stable Cas9 activity measured via GFP reporter assay¹⁰ in 1375 WT vs 226 *KRAS* mutant cancer cell lines, Wilcoxon rank-sum test P value 5.5E-05. (e) A scatter plot showing the difference in Cas9 activity between cell lines with a driver WT vs mutant for each driver (effect size, X-axis) and a corresponding Wilcoxon two-side significance for this difference (negative log₁₀-P value, Y-axis). (f) The change in mutant allele frequency (X-axis) of the *KRAS* mutations detected in different cell lines (cell line-mutation pair on the Y-axis) after induced Cas9 expression, compared to the corresponding parental cell lines, based on data from Enache *et al.*¹³. The starts and ends of arrows correspond to the mutant allele frequencies in the parental and the Cas9-expressed cell lines, respectively. Cases of increased allele frequency are colored in red, and those with

decreased frequency are colored in blue. (g) A scatter plot showing the median change in mutant allele frequency after induced Cas9 expression across all cell lines in Enache *et al.*¹³ (X-axis) and the corresponding Wilcoxon signed rank test significance (negative log₁₀-P value, Y-axis) for the 61 major cancer driver genes from Vogelstein *et al.*²⁷. The p-values are calculated using two-sided Wilcoxon Rank Sum tests unless not specified otherwise. In the boxplots of panels a, b and d, the center line, box edges and whiskers denotes the median, interquartile range and the rest of the distribution in respective order, except for points that were determined to be outliers using a method that is a function of the interquartile range, as in standard box plots.

KRAS mutant cells downregulate G2M checkpoint pathway in response to Cas9 induction

To investigate the mechanism underlying the potential selective advantage of *KRAS* mutant vs WT cells during CRISPR-KO, we analyzed gene expression data of 163 pairs of parental (without Cas9) and the corresponding Cas9-expressed cell lines (138 *KRAS* WT, 25 *KRAS* mutant)¹³, and identified the pathways that are differentially regulated upon Cas9 expression between *KRAS* WT and mutant cells (i.e. up/down-regulated in *KRAS* mutant but inversely or non-significantly regulated in *KRAS* WT cells; **Figure S9a, Supp. Note 5 & Figure S10** for *p53*). A major differentially regulated pathway was *G2M checkpoint* with the highest difference in the normalized enrichment score (**Figure S9b**), which is strongly downregulated (rank 2/50) in *KRAS* mutants but strongly upregulated in *KRAS* WT cells (4/50). This is a canonical pathway that serves to prevent the cells with genomic DNA damage from entering mitosis (M-phase) and thus its downregulation in *KRAS* mutant cells may provide them with a proliferative advantage [51]. Another top pathway, *E2F Targets*, which primarily regulates G1/S transition and DNA replication was also found to be downregulated in *KRAS* mutant cells but upregulated in *KRAS* WT cells upon Cas9 expression (**Figure S9b**). Thus, Cas9-induction may

similarly underlie the selective advantage of *KRAS* mutant cells by selectively activating cell cycle checkpoint pathways in response to DNA damage.

Discussion

In this study, we systematically investigated the possibility of selection of pre-existing cancer driver mutations during CRISPR-Cas9 gene editing. First, we confirmed and extended upon previous findings that selection^{11,12,13} of pre-existing *p53* mutations by CRISPR-Cas9 can happen, showing it in a large set of transformed and non-transformed cell lines. We identified the specific CDE+ genes whose CRISPR-KO is likely to mediate such selection, and further tested and validated some of these predictions in new screens and competitive assays that we have performed. After studying and validating our integrated computational and experimental pipeline in the known case of *p53*, we turned to apply it to study a collection of major cancer driver genes, and discovered that *KRAS* is another major cancer driver gene whose pre-existing mutants have a selective advantage during CRISPR-Cas9 gene editing. We demonstrated the selective advantage of *KRAS* mutant cells performing a CRISPR-KO/CRISPRi screen in isogenic cells with pooled CDE+ gene-targeting sgRNAs, and further in competition assays during the CRISPR-KO of top predicted *KRAS* CDE+ genes. We also observed that *KRAS* WT cells have lower Cas9 activity and thus a lower editing efficiency, similar to that observed for *p53*, which may limit CRISPR-mediated gene-editing in such cells²⁶. Analyzing recently published *KRAS* screens, we also find a subclonal expansion of *KRAS* mutant cancer cells following Cas9 expression. Finally, our study also shows that the introduction of the Cas9 protein downregulates the G2M checkpoint and E2F targets in *KRAS* mutant, but not *KRAS* WT cells, which may confer selective advantage to *KRAS*-mutant cells.

Multiple factors can contribute to the identity of CDE+ genes, including involvement in DNA repair and cell cycle pathways, being located in chromosomal fragile sites or highly accessible chromatin regions, supporting that their CRISPR-KO can lead to augmented DNA damage. We find that these factors can together account for up to 15% of our CDE+ genes. We also observed that a gene targeted by highly off-target guides can also lead to high DNA damage, which reassuringly only accounts up to 10% of the CDE+ genes (details in **Supp. Note 6**). Taken together, these three putative mechanisms can explain about 25% of the CDE+ genes we have identified, however the mechanisms underlying the rest are yet open to further studies.

Overall, our results point to a need for accounting for CDE effects in the analysis of dependencies in CRISPR screens. More importantly, our studies point to the need for careful selection of sgRNAs for therapeutic genome editing, and recommend cautionary monitoring of *KRAS* status in addition to that of *p53* during therapies utilizing CRISPR-Cas9. Lastly, in the publicly available CRISPR-Cas9 screens that we have analyzed, the current small numbers of cell lines with mutations in other cancer drivers, such as *VHL*, limits our ability to reliably determine whether these cancer genes could also be selected during CRISPR-Cas9 genome editing. The investigation of the latter thus awaits specifically designed screens in designated isogenic cell-lines.

Methods

CRISPR and shRNA essentiality screen data

We obtained CRISPR-Cas9 essentiality screen (or dependency profile) data in 436 cell lines from *Meyers et al.*¹⁰ for 16,368 genes, whose expression, CNV and mutation data are available via CCLE portal³³. We obtained the shRNA essentiality screen data in 501 cell-lines

from DepMap portal³⁴ for 16,165 genes, whose expression, CNV and mutation data is available publicly via CCLE portal³³. The 248 cell-lines and 14,718 genes that appear in both datasets were used in this analysis (**Table S1**). For mutation data, only non-synonymous mutations were considered. Synonymous (silent) mutations were removed from the pre-processed MAF files downloaded from CCLE portal³³.

Identifying CRISPR specific differentially essential genes of a potential CRISPR-selected cancer driver

For a given CCD (e.g. *p53* or *KRAS*), we checked which gene's essentiality (viability after knockout) is significantly associated with the mutational status of the CCD using a Wilcoxon rank sum test in the CRISPR and shRNA datasets, respectively (FDR<0.1). *CRISPR-specific differentially essential positive* (CDE+) genes are those whose CRISPR-KO is significantly more viable when the CCD is mutated while their shRNA silencing is not, whereas analogously CDE- genes are those whose CRISPR-KO is significantly more viable when the CCD is WT while their shRNA silencing is not. We filtered out any candidate CDE genes whose copy number was also significantly associated with the given mutation to control for potentially spurious associations coming from copy number (we removed genes showing significant association (FDR<0.1)) – the exact procedure used is described below in the section titled “*Identifying potential CRISPR-selected cancer drivers*”).

Identifying CDEs considering functional impact of mutations

Out of a total of 248 cell lines that we analyzed, 173 cell lines (69.7%) have *p53* non-synonymous mutations. In addition to identifying CDEs by considering all non-synonymous mutations, we additionally employed a more conservative approach where we aimed to consider

only *p53 loss-of-function* (LOF) mutations in the CDE identification process. To this end, we considered a mutation to be LOF if it was classified as non-sense, indel, frameshift, or among the 4 most frequent non-functional hotspot mutations (R248Q, R273H, R248W and R175H within the DNA-binding domain, determined as pathogenic by COSMIC³⁵). Using this definition we obtained new mutation profiles for *p53* and identified CDE genes via the same method described in the section titled “*Identifying CRISPR specific differentially essential genes of a potential CRISPR-selected cancer driver*”. We repeated a similar process with the top three known gain-of-function hotspot mutation variants of *KRAS*.

Identifying potential CRISPR-selected cancer drivers of CRISPR-KO

To identify additional CCD genes like *p53*, we considered 121 cancer driver genes identified by Vogelstein *et al.*²⁷, whose nonsynonymous mutation is observed in at least 10 cell lines (N=61). We determined whether each of these genes is a CCD as follows: for each of the 61 candidate genes, we tested the association between the essentiality of each of genes in the genome (reflected by post-KO cell viability) with the mutational status of the candidate CCD gene using a Wilcoxon rank sum test. We then counted the number of genes, whose essentiality is: (i) significantly positively associated with the candidate CCD mutational status (FDR-corrected p-value<0.1, median essentiality of WT>mutant of the cancer gene), (ii) significantly negatively associated with the candidate CCD mutational status (FDR-corrected p-value<0.1, median essentiality of WT<mutant of the cancer gene), and (iii) not associated (FDR-corrected p-value>0.1) with the candidate CCD mutation status; we performed this computation separately for the CRISPR and the shRNA screens, respectively. This computation results in a 3-by-2 contingency table for each candidate CCD gene. We then checked whether the distribution of the above three counts in the CRISPR dataset significantly deviates from that in the shRNA dataset

via a Fisher's exact test on the contingency table. If each of the values in the contingency table was greater than 30, we used the chi-squared approximation of the Fisher's exact test. We further filtered out any candidate CDE genes whose copy number was also significantly associated with the given mutation to control for potentially spurious associations coming from copy number (we removed genes showing significant association ($\text{FDR} < 0.1$)). We performed this procedure for all 61 candidate genes one by one and selected those with FDR corrected Fisher's exact test < 0.1 . We further filtered out the candidate CCD whose mutation profile is correlated with *p53* mutation profile via a pairwise Fisher test of independence ($\text{FDR} < 0.1$). We finally report the CCD genes that have a substantial number of CDE+ genes ($N > 300$).

Pathway enrichment analysis of CDE+/CDE- genes

We analyzed the CDE+/CDE- genes of each of the CCDs for their pathway enrichment with annotations from the Reactome database³⁶ in two different ways. First, we tested for significant overlap between our CDE genes with each of the pathways with hypergeometric tests ($\text{FDR} < 0.1$). Second, we ranked all the genes in the CRISPR-KO screen by the differences in their median post-KO cell viability values in mutant vs WT cells, and the standard GSEA method²¹ was employed to test whether the genes of each Reactome pathway have significantly higher or lower ranks vs the rest of the genes ($\text{FDR} < 0.1$). We repeated the GSEA analysis with the genes ranked by differential post-KD cell viability in the shRNA screen, and only reported significant pathways specific to CRISPR but not shRNA screens. We confirmed that for *p53*, the GSEA method was able to recover the top significant pathways identified by the hypergeometric test (e.g. those in **Figure 1d**), although extra significant pathways were identified (**Table S2**). For *p53* and *KRAS* CDE- genes respectively, the enriched pathways were clustered based on the

Jaccard index and the number of overlapping genes with Enrichment Map³⁷, and the largest clusters were visualized as network diagrams with Cytoscape³⁸.

To study the potential enrichment of CDE genes in common fragile sites (CFSs), we obtained chromosomal band locations of CFS¹⁶, and defined the CFS gene set as the set of all genes located within these chromosomal bands (obtained from Biomart³⁹). We tested for a significant overlap between our CDE genes and the CFS gene set with a hypergeometric test, and also confirmed the lack of significant overlap with the corresponding shRNA-DE genesets. Similarly, for the common highly accessible chromatin (HAC) regions, we obtained a list of these regions defined by a consensus of DNaseI and FAIRE across seven different cancer cell lines from a previous study⁴⁰. Next, we identified sgRNAs which are expected to target such HAC regions (see *Calculating off-target scores* section) and ranked genes based on the number of targeting such sgRNAs. Taking the top genes equal to the number of *p53* CDE+ genes, we computed the enrichment for *p53* CDE+ genes via a hypergeometric test.

Testing the clinical relevance of copy number alterations of the *p53* or *KRAS* CDE genes

We tested the hypothesis that copy number alterations in CDE+ genes (as a possible surrogate for the number of DSBs in these genes) can reduce the fitness of the CCD (*p53* or *KRAS*) WT tumors with patient data. The cancer genome atlas (TCGA)³² data of somatic copy number alteration (SCNA) and patient survival of 7,547 samples in 26 tumor types were downloaded from the UCSC Xena browser (<https://xenabrowser.net/>). In these tumor types *p53* is mutated in more than 5% of the samples. For each sample, the copy number alterations (genomic instability, GI) of a given set of genes, which quantifies the relative amplification or deletion of genes in a tumor based on SCNA was computed as follows⁴¹:

$$GI = \frac{1}{N} \sum_1^n I(s_i > 1)$$

where s_i is the absolute log ratio of SCNA of gene i in a sample relative to normal control, and $I()$ is the indicator function. Wilcoxon rank-sum test was then used to test whether the GI of CDE+ geneset is significantly lower than that of control non-CDE genes in CCD-WT but not in CCD-mutant tumors. Further, we tested if higher absolute levels of SCNA of the CDE+ genes are associated with increased rate of CCD (p53 or KRAS) mutation accumulation with cancer stage, as this would further testify that such amplification/deletion events in the CDE+ genes can drive the selection for CCD mutants. To this end, the following logistic regression model was used to identify the genes whose high absolute SCNA computed as above is associated with higher rate of CCD mutation accumulation with cancer stage, while controlling for cancer type and overall mutation load:

$$\text{logit}(P(\text{CCD})) = \beta_0 + \sum_k \beta_{\text{cancer_type}^k} \text{cancer_type}_k + \beta_{\text{mutation_load}} \text{mutation_load} + \beta_{GI} GI_i + \beta_{\text{stage}} \text{stage} + \beta_{\text{interact}} GI_i * \text{stage}$$

where CCD denotes the binary CCD mutational status of the patient, $\text{logit}(P(\text{CCD}))$ is the logit function of the probability of the CCD being mutant; cancer_type_k is the dummy variable for the category of the k^{th} cancer type; GI_i denotes the absolute value of SCNA levels of the given gene i as computed above; $GI_i * \text{stage}$ is the interaction term between the GI of gene i and cancer stage, that latter is made into a binary variable whose value is 0 for early stages (I and II) and 1 for late stages (III and IV). We tested the enrichment of CDE+ genes among the genes whose high absolute SCNA levels are significantly associated with higher rate of CCD mutation accumulation with cancer stage (i.e. genes with significantly positive β_{interact} coefficients in the above model) using a hypergeometric test.

Constructs and stable cell lines

MOLM13 cells were obtained from DSMZ (Cat. ACC-554) and maintained in RPMI-1640 medium (Life Technologies, Carlsbad, CA) supplemented with 10% v/v heat-inactivated fetal bovine serum (Sigma-Aldrich, Saint Louis, MI), 2 mM L-Glutamine (LifeTechnologies) and 100 U/mL penicillin/streptomycin (LifeTechnologies). *p53* R248Q was PCR amplified from a bacterial expression plasmid (kind gift of Dr. Shannon Lauberth, UCSD) and *KRAS*G12D the pBabe-*KRAS*G12D plasmid (Addgene plasmid 58902, from Dr. Channing Der) using the Kappa Hi-fidelity DNA polymerase (Kappa Biosystems). These PCR amplicons were separately cloned into the MSCV-IRES-tdTomato (pMIT) vector (a kind gift from Dr. Hasan Jumaa, Ulm) using Gibson Assembly. We first generated high-efficiency Cas9-editing MOLM13 leukemia cells by transducing these cells with the pLenti-Cas9-blasticidin construct (Addgene plasmid 52962 - from Dr. Feng Zhang) and selecting stable clones using flow-sorting. Clones were then tested for editing efficiency by performing TIDE analysis⁴². These MOLM13-Cas9 cells were then transduced retrovirally with the pMIT-*p53*R248Q or pMIT-*KRAS*G12D mutants and sorted for tdTomato using flow-cytometry (LSR Fortessa, BD Biosciences) to generate isogenic mutant MOLM13-Cas9 cell lines. Immortalized hTERT RPE1 cells were obtained from ATCC® (Cat. CRL-4000™) and maintained in DMEM-F12 medium (Life Technologies, Carlsbad, CA) supplemented with 10% v/v heat-inactivated fetal bovine serum (Sigma-Aldrich, Saint Louis, MI), 2 mM L-Glutamine (LifeTechnologies) and 100 U/mL penicillin/streptomycin (LifeTechnologies).

Generation of pooled sgRNA libraries

For pooled library cloning, 10 sgRNAs per gene were designed using the gene perturbation platform (<https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design>) Genetic Perturbation Platform. Guides targeting *p53* CDE+ and CDE- genes were synthesized as pools using array-based synthesis and cloned in the Lentiguide puro vector (Addgene plasmid 52963 - kind gift from Dr. Feng Zhang) using Golden Gate Assembly. In each assay, we have used ~240 unique non-targeting sgRNAs and 49 not expressing non-essential genes. A similar approach was used for the *KRAS* CDE libraries.

Pooled sgRNA library screen

30 million MOLM13-Cas9 cells or their isogenic MOLM13-*p53* or *KRAS* mutant counterparts were transduced with the pooled CDE library virus in RPMI medium supplemented with 10% fetal bovine serum, antibiotics and 8 µg/ml polybrene. The medium was changed 24 hours after transduction to remove the polybrene and cells were plated in fresh culture medium. 48 hours after transduction, puromycin was added at a concentration of 1 µg/ml to select for cells transduced with the sgRNA library. Puromycin was removed after 72 hours and then cells were cultured for up to 30 days. 7 days after transduction, approximately 4 million cells were collected, and genomic DNA was prepared for the time zero (T0) measurement and also from time 30 (T30). Genomic DNA from these cells was used for PCR amplification of sgRNAs and sequenced using a MiSeq system (Illumina). Fold depletion or enrichment of sgRNAs from the NGS data was calculated using PinAplPy software⁴³.

CDE+/- genes identified in isogenic experiments

From the read counts per million for each sgRNA at Day 0 and Day 30 from the above pooled CRISPR screens across two replicates, we removed all the sgRNAs with read count < 20 at Day 0. We calculated an average fold change (FC) of reads from Day 0 to Day 30. For each sgRNA, we calculated this FC-rank difference in *p53* WT vs mutant in both CRISPR-KO and CRISPRi screens. For consistent comparison with AVANA, we only considered sgRNAs used in both libraries. The top and bottom genes are *differentially essential (DE)* from each screen. Taking the top ranked genes based on the difference of this score in two screens, we identify the CDE+ and CDE- genes.

CRISPR Competition experiments

sgRNAs were cloned using standard cloning protocols and lentiviral supernatants were made from these sgRNAs in the 96-well arrayed format. 100,000 MOLM13 cells or tdTomato-positive isogenic mutants were plated in a 96 well plate and transduced with the sgRNA viral supernatants by spinfection with polybrene-supplemented medium. After selection of sgRNA transduced cells with puromycin for 48 hours, sgRNA transduced MOLM13 cells or mutants were mixed together in a ratio of 95:5 respectively, and the percentage of *p53* WT or *p53* mutant cells was monitored progressively up to 25 days using high-throughput flow-cytometry as described previously²³.

Quality control of publicly mined genetic screens used in the study

We first obtained gold-standard essential and non-essential geneset from Hart et al.⁴⁴. To test the quality of each genetic screen we computed an area under the precision-recall curve

(AUPRC) using the average logFC across replicates and cell lines. In this study, we only considered the genetic screens with an AUROC > 0.6 (random model AUPRC=0.5). We also employed this method to test the quality of our in-house generated genetic screens.

CRISPR-Cas9 Ribonucleoprotein transfection experiments

We generated sgRNAs by *in vitro* transcription using the HiScribe™ T7 Quick High Yield RNA Synthesis Kit (New England Biolabs, Beverly, MA) and performed the Ribonucleoprotein (RNP) complex formation using TrueCut Cas9 Protein v2 (ThermoFisher Scientific, Waltham, MA) according to published protocols⁴⁵. MOLM13 cells without Cas9 and expressing pMIT-*p53R248Q* or pMIT-*KRAS* were generated as described in the *Constructs and stable cell lines* section. 1M cells were transfected with NFDUFB6 sgRNA or NTC sgRNA in triplicates with 1mg of Cas9 and 1mg of RNA in 10 ml of Buffer R using the Neon™ transfection system (ThermoFisher Scientific; 1500V, 20 ms, single pulse). Cells were maintained in culture for 48 hours before harvest for imaging, dye-dilution and editing estimation assays. For *NDUFB6* and NTC editing estimation, we used the Synthego Performance Analysis ICE tool according to the instructions, using un-transfected parental MOLM13 samples as controls and samples from 48 hours post-transfection as the Day 0 initial timepoint and Day 10 as a final time point, in triplicates. For RPE1 experiments, mutant cells with *p53R248Q* or pMIT-*KRAS* similarly and transfections were performed using Lipofectamine™ CRISPRMAX™ Cas9 Transfection Reagent (ThermoFisher Scientific) according to the manufacturer's instructions for 12-well plate format, in triplicates. Cell harvesting time-points were similar to those of MOLM-13.

Dye-dilution experiments

We used the CellTrace™ Violet Cell Proliferation Kit (ThermoFisher Scientific) to stain MOLM13-WT and MOLM13-p53 mutant cells transfected with Cas9 RNP complexed with NTC or NDUFB6 RNA, according to the manufacturer's instructions. Cells were maintained in culture in the dark and assayed by flow cytometry using the LSR Fortessa every 2 days for 14 days. FCS files were analyzed using FlowJo software.

Analysis of γ -H2AX foci in MOLM13-Cas9 and MOLM13-p53 mutant cells

MOLM13-WT and MOLM13-p53 mutant cells were left untreated or treated with 1 μ M doxorubicin for 2 h at 37°C 5% CO₂, which served as negative and positive controls for DNA damage mediated γ -H2AX foci formation, respectively. MOLM13-WT and MOLM13-p53 mutant cells transfected with Cas9 RNP complexed with NTC or NDUFB6, and negative control cells were pelleted at 400g for 5 min at 4°C, washed two times in PBS and fixed in 4% paraformaldehyde in PBS for overnight at 4°C. The cells were washed two times in PBS and permeabilized in 0.25% triton X-100 in PBS for 5 min at room temperature. Following two washes with PBS, the cells were incubated in blocking buffer (3% BSA in PBS) for 30 min at room temperature and subsequently incubated with APC conjugated H2AX phospho (Serine 139) antibody (BioLegend; Cat # 613415) at an antibody dilution of 1:200 in blocking buffer for overnight at 4°C in dark. Cells were washed two times with PBS and resuspended in 150 μ l PBS. Cell suspensions were spotted on poly-lysine coated glass slides using cytopsin (Cytospin 4; Thermo Scientific) centrifugation at 800 rpm for 4 min. Coverslips were mounted onto the slides using ProLong Gold antifade reagent with DAPI (Invitrogen) and cured for overnight at room temperature in dark. Slides were imaged in Nikon A1R HD confocal microscope. Sequential z-

sections were imaged using a 60x oil objective and maximum projection images were obtained using the Nikon NIS-Elements platform.

Cas9 activity in cancer cell lines with KRAS (or other cancer driver) WT vs mutant

We downloaded the exogenous Cas9 activity of 1601 cancer cell lines from DepMap portal and their KRAS mutation status considering only non-synonymous variants profiled using whole exome sequencing (1375 and 226 *KRAS* WT and mutant, respectively)¹⁵. We tested whether the Cas9 activity is higher in *KRAS* mutant vs *KRAS* WT cell lines using one-sided wilcoxon rank-sum test. We repeated this process for each cancer driver gene and used the FDR corrected significance to rank them in addition to the fold change of Cas9 expression.

Subclonal expansion of KRAS mutant in parent vs high Cas9-expressed cell lines

We downloaded the deep targeted sequencing of cancer driver genes performed on 42 parental and matched Cas9-expressed cancer cell lines from Enache et al.¹³. In this analysis, we discarded the cell lines with <20% Cas9 activity and thus low DNA damage. We asked whether mutant allele frequency of a cancer driver (e.g. *KRAS*) significantly increased in Cas9-expressed cell lines compared to matched parental cell lines using Wilcoxon signed-rank test. In this analysis, we have considered both intronic and exonic variants provided from sequencing.

Analysis of differentially expressed pathways in KRAS wildtype and mutant cells in response to Cas9 induction

Gene expression profiles of 163 pairs of parental (without Cas9) and the corresponding Cas9-expressed cell lines (138 *KRAS* WT, 25 *KRAS* mutant) were obtained from Enache et al.¹³. Differential expression analysis between the Cas9-expressed cells and the parental cells was

performed for the *KRAS* WT and mutant cells separately, and GSEA analysis²¹ (genes ranked by logFC) was performed to identify the hallmark pathways from MSigDB⁴⁶. We next identified pathways that are differentially regulated upon Cas9 expression between *KRAS* WT and mutant cells. These include the pathways that are up-regulated in the *KRAS* mutant cells but down-regulated or non-significantly altered in the *KRAS* WT cells, and vice versa. The pathways are ranked by the difference of normalized enrichment score in WT vs mutant cells. This analysis is performed using the *fgsea* R package²¹.

Data and Code availability Statement

We have provided the scripts and data from both previously published and in-house screens, in their raw and processed form to reproduce each step of results and plots in a GitHub repository which can be accessed here: https://github.com/ruppinlab/crispr_risk

Acknowledgements

We acknowledge and thank the National Cancer Institute for providing financial and infrastructural support. We thank Curtis Harris, Andre Nussenzweig, Sridhar Hannenhalli and the members of Cancer Data Science Lab for insightful feedback. This research was supported in part by the Intramural Research Program of the National Institutes of Health, NCI. S.S and K.C. are supported by the NCI-UMD Partnership for Integrative Cancer Research Program. AJD would like to acknowledge the support of the National Cancer Institute of the National Institutes of Health under Award Number P30 CA030199, the Rally Foundation for Childhood Cancer Research and Luke Tatsu Johnson Foundation under Award Number 19YIN45, an Emerging

Scientist Award from the Children's Cancer Research Fund, and the V Foundation for Cancer Research (TVF) under Award Number DVP2019-015.

References

1. Hsu, P. D., Lander, E. S. & Zhang, F. Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell* **157**, 1262–1278 (2014).
2. Kim, D. *et al.* Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods* **12**, 237–43, 1 p following 243 (2015).
3. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
4. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
5. Cullot, G. *et al.* CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. *Nat. Commun.* **10**, 1136 (2019).
6. Charlesworth, C. T. *et al.* Identification of Pre-Existing Adaptive Immunity to Cas9 Proteins in Humans. doi:10.1101/243345
7. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
8. Aguirre, A. J. *et al.* Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov.* **6**, 914–929 (2016).

9. Munoz, D. M. *et al.* CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. *Cancer Discov.* **6**, 900–913 (2016).
10. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics* **49**, 1779–1784 (2017).
11. Ihry, R. J. *et al.* p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. *Nat. Med.* **24**, 939–946 (2018).
12. Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. *Nature Medicine* **24**, 927–930 (2018).
13. Enache, Oana M., et al. "Cas9 activates the p53 pathway and selects for p53-inactivating mutations." *Nature Genetics* (2020): 1-7.
14. Schirotti, Giulia, et al. "Precise gene editing preserves hematopoietic stem cell function following transient p53-mediated DNA damage response." *Cell Stem Cell* 24.4 (2019): 551-565.
15. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–576.e16 (2017).
16. Lukusa, T. & Fryns, J. P. Human chromosome fragility. *Biochim. Biophys. Acta* **1779**, 3–16 (2008).
17. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nature Biotechnology* **36**, 765–771 (2018).
18. van den Berg, J. *et al.* DNA end-resection in highly accessible chromatin produces a toxic

break. doi:10.1101/691857

19. Richardson, C. D. *et al.* CRISPR–Cas9 genome editing in human cells occurs via the Fanconi anemia pathway. *Nature Genetics* **50**, 1132–1139 (2018).
20. Behan, Fiona M., et al. "Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens." *Nature* 568.7753 (2019): 511.
21. Sergushichev, Alexey A. "An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation." *BioRxiv* (2016): 060012.
22. Brown, Kevin R., et al. "CRISPR screens are feasible in TP 53 wild-type cells." *Molecular systems biology* 15.8 (2019): e8679.
23. Deshpande, A. *et al.* Investigation of Genetic Dependencies Using CRISPR-Cas9-based Competition Assays. *J. Vis. Exp.* (2019). doi:10.3791/58710
24. Boettcher, Steffen, et al. "A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies." *Science* 365.6453 (2019): 599-604.
25. Giacomelli, Andrew O., et al. "Mutational processes shape the landscape of TP53 mutations in human cancer." *Nature genetics* 50.10 (2018): 1381
26. Tario, Joseph D., et al. "Monitoring cell proliferation by dye dilution: considerations for probe selection." *Flow Cytometry Protocols*. Humana Press, New York, NY, 2018. 249-299.
27. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546–1558 (2013).
28. Hähnel, P. S. *et al.* Targeting components of the alternative NHEJ pathway sensitizes KRAS mutant leukemic cells to chemotherapy. *Blood* **123**, 2355–2366 (2014).

29. Jinesh, G. G., Sambandam, V., Vijayaraghavan, S., Balaji, K. & Mukherjee, S. Molecular genetics and cellular events of K-Ras-driven tumorigenesis. *Oncogene* **37**, 839–846 (2018).
30. Luo, Ji, et al. "A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene." *Cell* 137.5 (2009): 835-848.
31. Martin, T. D. *et al.* A Role for Mitochondrial Translation in Promotion of Viability in K-Ras Mutant Cells. *Cell Rep.* **20**, 427–438 (2017).
32. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45**, 1113–1120 (2013).
33. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
34. McFarland, J. M. *et al.* Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun.* **9**, 4610 (2018).
35. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* **43**, D805–D811 (2015).
36. Matthews, L. *et al.* Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37**, D619–22 (2009).
37. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).
38. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular

interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

39. Smedley, D. *et al.* The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* **43**, W589–98 (2015).

40. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* **21**, 1757–1767 (2011).

41. Bilal, E. *et al.* Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS Comput. Biol.* **9**, e1003047 (2013).

42. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168 (2014).

43. Spahn, P. N. *et al.* PinAPL-Py: A comprehensive web-application for the analysis of CRISPR/Cas9 screens. *Sci. Rep.* **7**, 15854 (2017).

44. Hart, Traver, *et al.* Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Molecular systems biology* 10.7 (2014).

45. Brunetti, L., Gundry, M.C., Kitano, A., Nakada, D., Goodell, M.A.. Highly Efficient Gene Disruption of Murine and Human Hematopoietic Progenitor Cells by CRISPR/Cas9. *J Vis Exp.* **134**, 57278 (2018).

46. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., Tamayo, P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **6**, 417–425 (2015).

Chapter 3: Designing optimal combination treatment targeting the clonal architecture of the tumor using scRNA-seq

Abstract

A combination treatment designed to target the multiple clones of a patient's tumor could decrease the likelihood of resistant emergence. The ability to learn and predict the response of a drug at a single-cell resolution could help to design such optimal combinations. A lack of large-scale patients' datasets with single-cell (sc) expression hinders our ability to build such models. To overcome this limitation, we built a precision oncology framework for personalized single-cell expression-based planning for treatments in oncology (PERCEPTION) that utilizes large-scale drug screens in cancer cell lines and matched bulk and single-cell transcriptome profiles to build response models that can be translated to patients. We first showed that our predicted viability profile of multiple drugs with known mechanisms of action strongly correlates with the targeted pathway activity at a single-cell resolution, demonstrating our ability to predict at this resolution. We next predict response to monotherapy and combination treatment in three independent screens performed across cancer cell lines and patient-tumor-derived cell lines using their sc-expression profiles. Translating to the clinical context, we successfully stratify responders of combination therapy in a multiple myeloma clinical trial using the tumor's sc-expression profile. For individuals in this trial, we next find the optimal combination treatments (pairs and triplets) from the existing FDA-approved drugs set, where the pair of gefitinib and niraparib, an EGFR and a PARP inhibitor, is one of the top candidates. In sum, we provide a first-of-its-kind framework to utilize the tumor's sc-expression to identify responders to combination therapy or to design an optimal combination for an individual patient.

Introduction

Tumors are typically heterogeneous and composed of numerous different clones, making treatments targeting multiple clones more likely to diminish the likelihood of resistance emerging due to clonal selection, enhancing the overall patient's response (Castro et al. 2021). Bearing this goal in mind, large-scale combinatorial pharmacological screens in patient-derived cell lines, xenografts, and organoids have given rise to numerous combination treatment candidates (Wensink, et al. 2021, Yao et al. 2020, de Witte et al. 2020). However, these studies are limited by the large number of combinations that are needed to be tested in various genomic contexts and thus, there is a need for *in silico* methods to narrow down the search space.

The characterization of the tumor microenvironment via single-cell omics has already led to countless important insights regarding the complex network of tumor-immune interactions involving many different cell types (Castro et al. 2021). It also offers a promising way to learn and predict drug response at a single-cell resolution. The latter, if successful, could guide the design of drug combinations that target multiple tumor clones disjointly (Shalek et al 2017, Adam et al 2020 & Zhu et al 2017). However, building such predictors of drug response at a single cell (SC) resolution is very challenging due to the paucity of large-scale preclinical or clinical training datasets. Previous efforts have been primarily restricted to pre-clinical models in just a few cell lines (Kim et al 2016, Suphavitai et al 2020). Yet, efforts to identify biomarkers of response and resistance at the patient level using single-cell expression are rapidly emerging for both targeted- and immuno- therapies (Cohen et al 2021, Ledergor et al 2018, Sade-Feldman et al 2018).

Aiming to address this challenge systematically, here we present a precision oncology framework for PERsonalized single-Cell Expression-based Planning for Treatments In ONcology (PERCEPTION) that builds upon the recent availability of large-scale pharmacological screens and SC expression data in cancer cell lines to construct machine learning based predictors of SC drug response. First, using these predictors, we show that the predicted viability for drugs with known mechanisms of action correlates with the pathway activity it is targeting at a single-cell resolution. Second, we show that the SC-based models can successfully predict the response to single and combination treatments in three independent screens performed in cancer and patient-tumor-derived cell lines based on their SC-expression profiles. Thirdly, we show that SC-based models successfully stratify responders to combination therapy in a multiple myeloma clinical trial based on their tumor's SC-expression data. Finally, we identify combination treatments (pairs and triplets) of existing FDA-approved oncology drugs that kill tumor clones as effectively as possible.

Results

Overview of PERCEPTION

To predict patients' response to a therapy using their tumor's single-cell SC-expression profile, we built a machine learning pipeline called *PERCEPTION* (**Figure 1A**, detailed description is provided in **Methods**). *PERCEPTION* builds drug response models from large-scale pharmacological screens performed in cancer cell lines where bulk and SC-expression are available. As there is currently a paucity of large-scale matched response and single-cell data either in pre-clinical or patients, we designed a prediction pipeline that first is trained on large-scale bulk-expression profiles of cancer cell lines and then, in a second step, its performance is

further optimized by tuning the available SC-expression profiles of cancer cell lines. To this end, we mined bulk-expression (Ghandi et al 2020) and drug response profiles (PRISM) of 100s of different cancer cell lines (N=488, **Table S1**) from the DepMap database (Tsherniak. et al 2017). The SC-expression profiles of these cell lines (N=205, **Table S1**) are obtained from Kinker et al. 2020. Drug efficacy is measured via area under the curve (AUC) viability-dosage curve, where lower AUC values indicate increased sensitivity to treatment (**Table S1**). Briefly, for a given drug X, *PERCEPTION* performs the following two steps: **1.** It first builds a regularized linear response prediction model from the bulk expression and drug response data available for ~300 cancer cell lines. **2.** In the second step, the hyperparameters of *PERCEPTION* were tuned further to maximize its ability to predict the response from SC-expression data i.e. ~160 cancer cell lines with matched SC-expression and response to the drug. The output of our pipeline is a response model and a quantification of its predictive accuracy from SC-expression in never seen before leave-one-out test data.

Illustration of *PERCEPTION*'s ability to predict viability at single-cell resolution via two case studies

To visualize *PERCEPTION*'s ability to predict cell killing at single-cell resolution, we examined our predicted killing for two drugs, where the mechanism of action pathway of the drug is well characterized (Nutlin-3 and Erlotinib). We applied the *PERCEPTION* pipeline described above to build SC-based predictors for these two drugs and studied them further, as follows.

The first case involves the canonical antagonist, Nutlin-3, whose mechanism of killing involves the inhibition of the interaction between MDM2 and tumor suppressor p53; thus,

MDM2 high activity is a known response biomarker to nutlin-3 treatment (Arya et al. 2010). Via *PERCEPTION*, we built a response model for Nutlin-3, whose correlation with the observed response at the bulk expression was 0.598, $P=1.2E-16$, (with *MDM2* expression being one top-ranked predictive features). We predicted the killing post its treatment for 3566 single cells across nine p53 wild-type lung cancer cell lines. Across these single-cells, we observed that that the predicted killing after nutlin-3 treatment and MDM2 expression are strongly correlated across the individual cells screened (Pearson Rho= 0.50, $P<2E-16$, as visualized in **Figure 1D**), as expected. We also find that we can identify sub-clones with pre-existing nutlin-3 resistance (**Figure 1D-arrow highlight**). In the second case, we repeated this analysis to study and visualize *PERCEPTION*'s ability to predict the response to erlotinib, which targets the gain-of-function mutation in epidermal growth factor (EGFR) and is mainly used to treat lung cancer patients with activating EGFR mutations. We found that individual cells with low EGFR pathway activity signature are predicted to be resistant to this treatment, as expected, and that the predicted and observed killing levels are correlated across individual cells (Pearson Rho= 0.42, $P<2E-16$) as visualized in **Figure 1E**. Analogous findings for other EGFR inhibitors developed more recently than erlotinib are provided in **Extended Figure 1B**.

Testing *PERCEPTION* predictions for FDA approved drugs

We applied *PERCEPTION* to build SC predictors of response for 133 U.S FDA-approved oncology drugs available in the drug screen (PRISM) (**Table S2**). The predictive performances for these drugs are provided in **Figure 1B**. We defined models as predictive if the Spearman correlation between their predicted vs observed viability was greater than 0.3. This threshold was chosen as it corresponds to the mean cross-screen replicate correlation observed among three major pharmacological screens (average cross-platform correlation across GDSC, CTD &

PRISM ~ 0.30 (Corsello et al. 2020)). We were able to build models for 33% (44 out of 133 drugs, **Table S2**) of the total drugs tested whose prediction accuracy exceeds this threshold (**Figure 1B**). Studying the predictive accuracy of these 44 predictive models in a cross-validation manner for different transcriptomics inputs, including SC, bulk, and pseudo-bulk-expression (generated by summing up the gene-mapped reads across single cells, Methods), reassuringly we find that the predictive performance based on SC-expression is comparable to that obtained using bulk-expression or pseudo-bulk (**Figure 1C**).

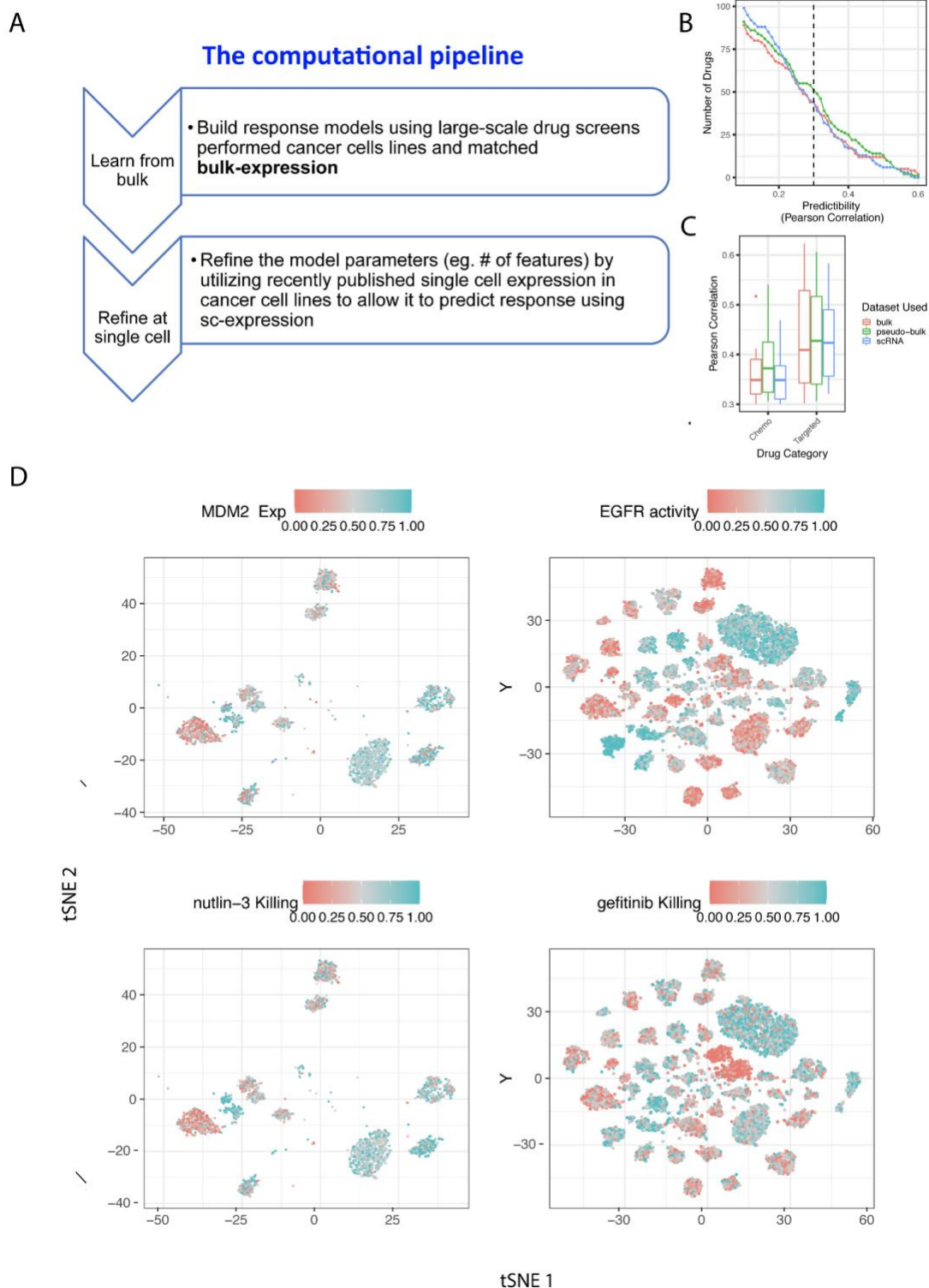


Figure 1. PERCEPTION-based precision oncology framework. (A) The PERCEPTION precision oncology framework is composed of two steps: (i) Build response models using

large-scale drug screens performed on cancer cell lines and matched bulk-expression.

(ii) Refine the model parameters i.e., number of features by utilizing sc-expression in cancer cell lines to enable our model for the usage of sc-expression. **(B)** The number of predictive models for FDA-approved drugs for cancer generated by PERCEPTION during cross-validation (y-axis) when sc-expression (blue), bulk-expression (red), and pseudo-bulk are used for a Pearson correlation threshold (x-axis, Predictability) **(C)** The distribution of predictive performance (x-axis) of drugs with a predictive model, defined by Pearson $Rho > 0.3$. In the boxplots, the center line, box edges, and whiskers denote the median, interquartile range, and the rest of the distribution, respectively, except for points that were determined to be outliers using a method that is a function of the interquartile range, as in standard box plots. **(D)** In the left-most panel, killing a canonical MDM2 antagonist, Nutlin-3 and expression of MDM2 are provided for every single cell (each point) in the top and bottom tSNE plot, respectively. The intensity of the color denotes the extent of killing in the left panel and MDM2 expression in the right pane, where the respective legends are provided. In this panel, we provide 3566 single-cells from nine p53 WT lung cancer cell lines. The tSNE clustering is performed using the expression profile of all the genes.

PERCEPTION predictive performance in an independent large-scale GDSC drug screen

We next tested the performance of our models on an independent screen, GDSC (Garnett et al. 2012). To this end, we first identified drugs shared between the PRISM and GDSC screens (N=191, **Table S3**, quality control and model building steps in **Methods**). We were able to build PRISM-based predictive models for 16 of these common drugs. The mean correlation between experimental viability reported in GDSC vs PRISM (screen

concordance) across 80 shared cell lines serving as an unseen testing set was 0.44. For the same testing set, the mean correlation between the predicted vs observed viability in 0.38 and 0.28 in PSISM and GDSC, respectively (**Figure 2A**). As expected, the prediction performance of a model in the GDSC test set is correlated with the concordance between the drug's viability profiles in GDSC and PRISM datasets (Pearson Rho=0.49, P=5.89E-02; **Figure 2B, Table S4**). As the range of predicted values is smaller than observed (**Extended Figure 2**), we use a scaled predicted AUC (z-score) in further analyses reported below.

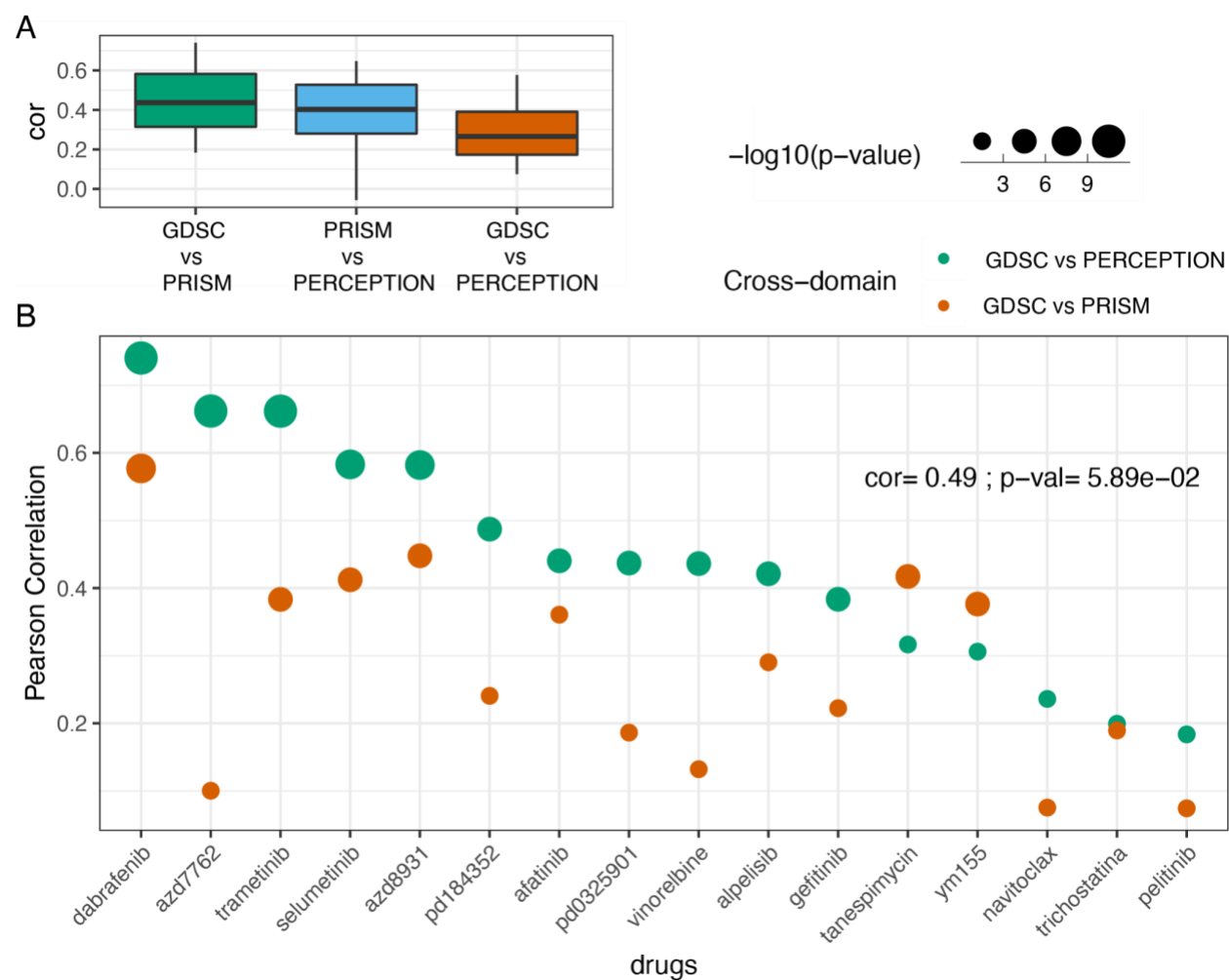


Figure 2: PERCEPTION's performance in the GDSC screen. (A) The correlation

measure via Pearson Rho (y-axis) comparing “GDSC vs PRISM”, “PRISM vs PERCEPTION”, and “GDSC vs PERCEPTION”. Drug response predictions were performed at a single-cell resolution and the cell line level response (mean response across single cells) was used to compare. **(B)** Relationship between the concordance in “GDSC vs PRISM” (green) and “GDSC vs PERCEPTION” (orange) across 16 drugs. These are the drugs with pharmacological screens in both PRISM and GDSC platforms and that have a substantial positive correlation between their AUC values ($cor > 0.3$ and $p\text{-value} < 0.5$ in cell lines excluding the above 80 cell lines). The size of the dots represents the Pearson correlation-based p-value in $-\log_{10}$ scale. The drugs are ordered on the x-axis from left to right in the decreasing order of their correlation between GDSC and PRISM responses.

SC-based PERCEPTION models prediction of monotherapy and combination response in a lung cancer cell lines screen

To study *PERCEPTION* in another independent screen, we tested its predictive performance in a recent drug screen in lung cancer cell lines (Nair et al. 2021). We focused on cancer 40 drugs that are FDA-approved or in clinical trials for which we could build predictive *PERCEPTION* models. We assessed their predictive performance vs drug screen data measured for monotherapy and two-drug combinations of 14 of these drugs (whose viability profiles passed quality control) across 21 lung cancer cell lines in five dosages (**Table S5, methods, Supp Note 1**). Matched SC-expression was mined for these lung cancer cell lines from Kinker et al. 2020, including about 300 cells per cell line.

Given this data, we used the *PERCEPTION* models to predict the response to each drug in each cell line by computing the mean predicted viability across all the single cells of that from a cell line. The predicted viability is significantly higher in resistant vs sensitive cell lines (**Figure 3A**, top vs bottom 33% cell lines ranked by viability, Wilcoxon rank-sum $P=2E-06$, $FC=1.53$), and can stratify responders vs non-responders (ROC-AUC = 0.72, **Figure 3B**). The predictive performance of high-confidence screen results (see **Methods**) is considerably higher (AUC=0.88, **Figure 3C-blue curve**, $FC=1.95$, **Figure 3B-right panel**). The overall mean Pearson correlation between predicted vs observed viability for these drugs is 0.33; $P<7.4E-09$ (**Extended Figure 3A**). Detailed drug level comparisons between observed vs predicted viability are provided in **Extended Figure 3B**.

We next tested *PERCEPTION*'s ability to predict the response to 42 combinations of these 14 drugs studied in this screen (**Table S5**). A combination response in a given cell line was predicted by adopting the independent drug action (IDA) model across all the single cells from that cell line (Ling et al. 2020) i.e. the combination response of two drugs is simply the effect of the single most effective drug in the combination (**Supp Note 2**). The predicted combination viability is significantly higher in resistant vs sensitive cell lines (Wilcoxon rank-sum $P=8.3E-03$, $FC=1.54$, **Figure 3D**) and can stratify the responders vs nonresponders (AUC=0.69, # of resistant data points=28, # of sensitive data points=24, **Figure 3E**). Like in the monotherapies case, the observed effect size is considerably higher when considering only high-confidence screen results ($P<8.8E-03$, $FC=1.77$, **Figure 3D-right panel**, AUC=0.87, **Figure 3E-blue curve**, Pair-level comparison between observed vs predicted are provided in **Extended Figure 3D**).

Taken together, these results to the ability of *PERCEPTION* models to predict single and combination therapies in independent screens without any further training.

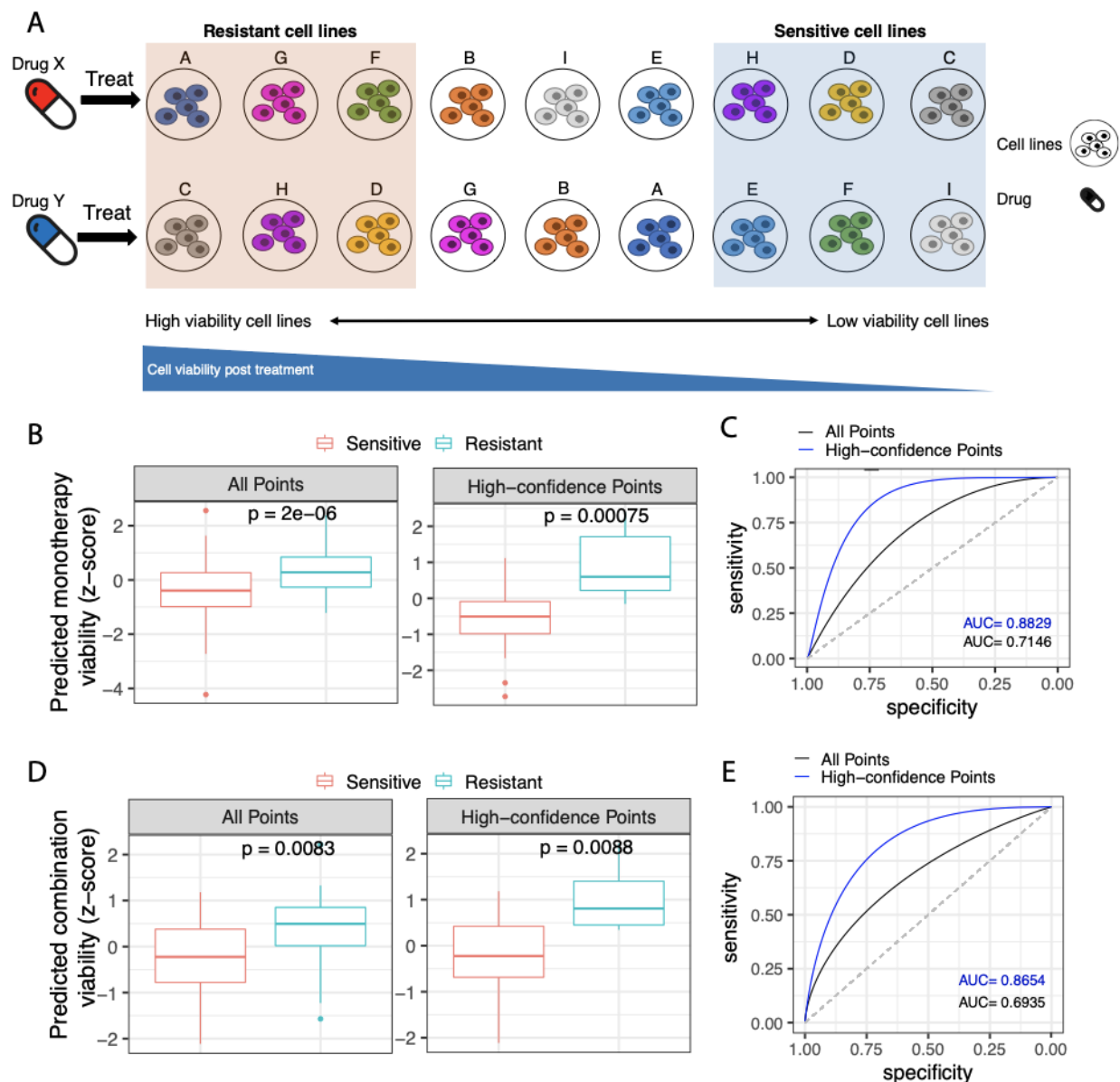


Figure 3: PERCEPTION's predicted response successfully stratifies *resistant* vs *sensitive* lung cancer cell lines to monotherapy and combination response. A) Illustration of our method defining sensitive and resistant cell lines for treatment from viability profile, where for each drug

the top and bottom 33% of the cell lines ranked by viability in ascending order are defined as sensitive and resistant, respectively. **B)** The predicted viability from *PERCEPTION* (x-axis) for top vs bottom 33 % cell lines defined as resistant (N=72) vs sensitive (N=84) cell lines, considering all the data points (left panel) or only high-confidence points (right panel) via a standard boxplot. A respective significance computed using a one-tailed Wilcoxon rank-sum test is provided. **C)** A receiving operator curve is plotted showing the relationship between sensitivity and specificity, where the area under the curve denotes the power of stratification. The colors of the curves - black and blue, represent whether all data points or only high-confidence data points are used, respectively. The area under this curve is provided at the right corner and similarly color-coded. The area under the dashed diagonal line denotes a random-model performance. Like **B)** and **C)**, panel **D)** and **E)**, respectively shows our ability to predict combination viability (# of resistant data points=28, # of sensitive data points=24).

SC-based *PERCEPTION* prediction in patient-derived Head and Neck cancer cell lines

To test the ability of our models to predict response in patient-derived cell lines (PDC), we used SC-expression of head and neck cancer cell lines derived from five different patients treated with eight different drugs at two concentrations (**Table S6**), with both monotherapy and combination therapy (Suphavilai et al. 2020). We were able to build predictive *PERCEPTION* response models for 4 out of the 8 drugs (docetaxel, epothilone-b, gefitinib, and vorinostat; Pearson Rho threshold > 0.25). For monotherapy treatments, the predicted viability is significantly higher in resistant vs sensitive cell lines (N=16 each, **Figure 4A**), with an AUC of 0.64 (**Figure 4B**). The predicted viability is correlated with the observed viability (Pearson Rho=0.46; P<0.03, **Extended Figure 5A**), and individual drug-level correlations are provided in **Extended Figure 4C**. Higher predicted viability in resistant cell lines is also observed for

combination treatments (**Figure 4C**), with an AUC of 0.86 (**Figure 4D**). The predicted viability after gefitinib treatment is illustrated at a single-cell resolution in **Figure 4E**. The predicted vs experimental correlations obtained for all data points and drug levels are provided in **Extended Figure 4B, D**. These results demonstrate the ability of *PERCEPTION* models to predict response in patient-derived single cells.

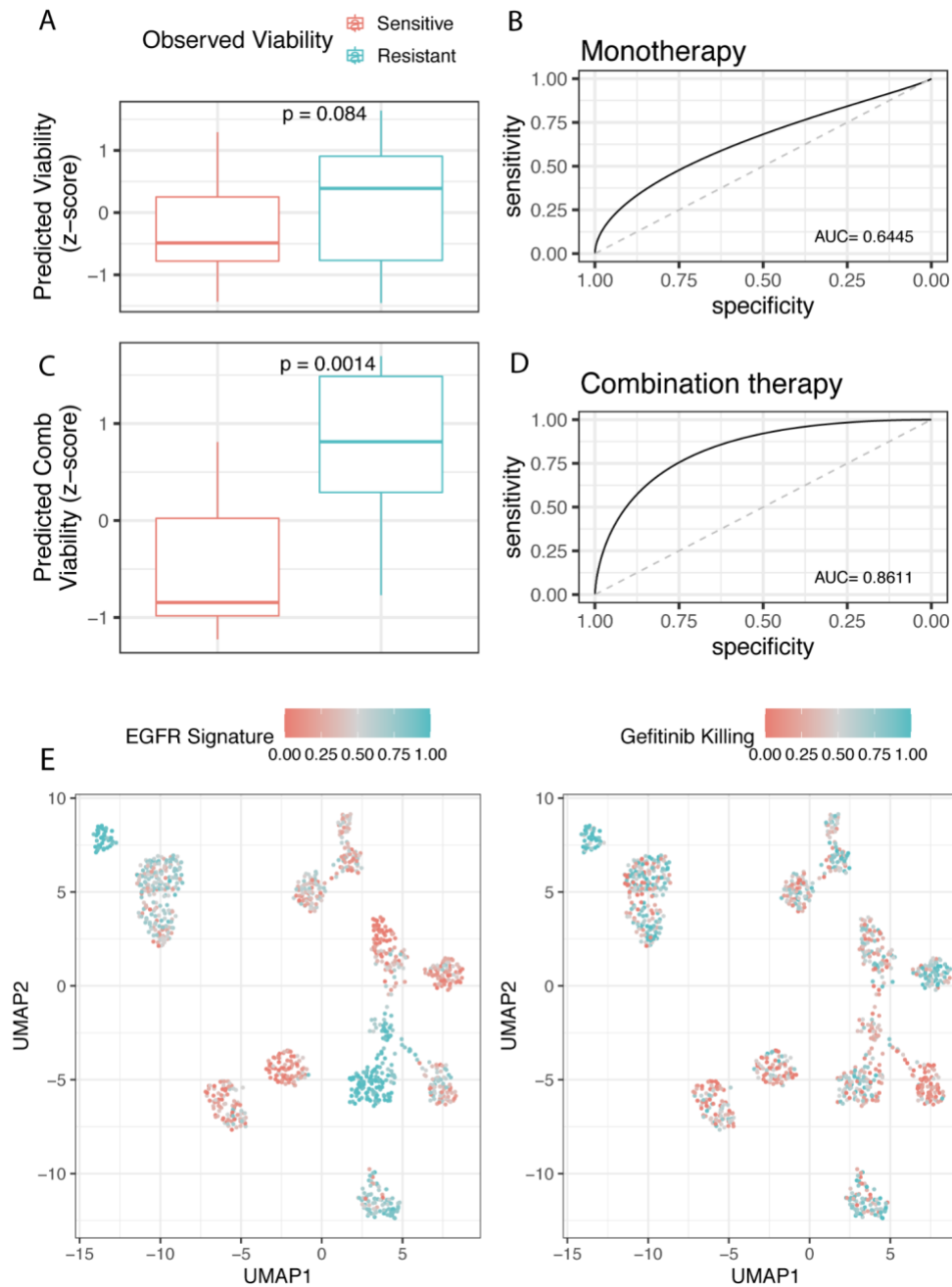


Figure 4: Prediction monotherapy and combination response in patient-derived cell lines.

(A) The predicted viability from *PERCEPTION* in resistant (n=16) vs sensitive (n=16) cell lines.

(B) ROC curves depicting the prediction power (sensitivity and specificity) of the predicted viability to stratify resistant vs sensitive cell lines. The area under this curve is provided at the right corner and denotes overall model prediction power. The area under the dashed diagonal line denotes a random-model performance. In **(C)** and **(D)**, we repeated the analysis for combination treatment (Number of resistant vs sensitive cell lines=12 vs 12). The boxplots provided are standard and the significances are computed using the one-tailed Wilcoxon rank-sum test. **E)** Viability after treatment with gefitinib, a canonical EGFR mutation inhibitor, and activity of the EGFR pathway is provided for every single cell (each point) in the left and right UMAP plot, respectively. The intensity of the dot color denotes the EGFR pathway activity in the left panel and viability in the right panel, where the respective legends are provided. In this panel, we analyze 1116 single-cells from 5 PDCs. The UMAP clustering is performed using the expression profile of all the genes.

***PERCEPTION* Prediction of combination treatment response in a Multiple myeloma clinical trial**

We turn to test the ability of *PERCEPTION* models to predict patients' responses based on SC transcriptomics from their tumors, our main goal. Performing a literature search for clinical trials of targeted or chemotherapy that report both patient's tumor SC-expression data and response labels, we found one such dataset with 41 multiple myeloma patients. The cells were clustered in the original paper to three clones (median) based on their expression profiles. These patients were treated with a DARA–KRd combination of four drugs - daratumumab,

carfilzomib, lenalidomide, and dexamethasone (Cohen et al. 2021). SC-expression and response labels were available for 28 of these patients, whose pretreatment sub-clonal distribution is shown in **Figure 5A**. Patient response was measured via tumor size estimates in radiological images.

We built *PERCEPTION* response models for two out of four of these drugs (carfilzomib and lenalidomide) whose response profiles are available in either PRISM or CTD. Using these models, we predicted a combination response in a given patient via the following two steps. We first predicted the combination response for each clone of a patient using their average expression profile. We observed that the response observed in the most resistant clone of a patient best stratifies the responders vs non-responder patients ($P < 1.9 \times 10^{-3}$, **Figure 5C**).

Therefore, the predicted response of a patient is the response of the most resistant clone available (**Methods**). The resulting predicted response of a patient can successfully stratify the responders vs non-responders with an AUC of 0.827 (**Figure 5D**). In comparison, repeating the above analysis using pseudo-bulk expression (computed here as a mean expression over all the cells in the tumor) yields an AUC of 0.56, testifying to the marked benefit of harnessing SC data from patients' tumors to predict their response.

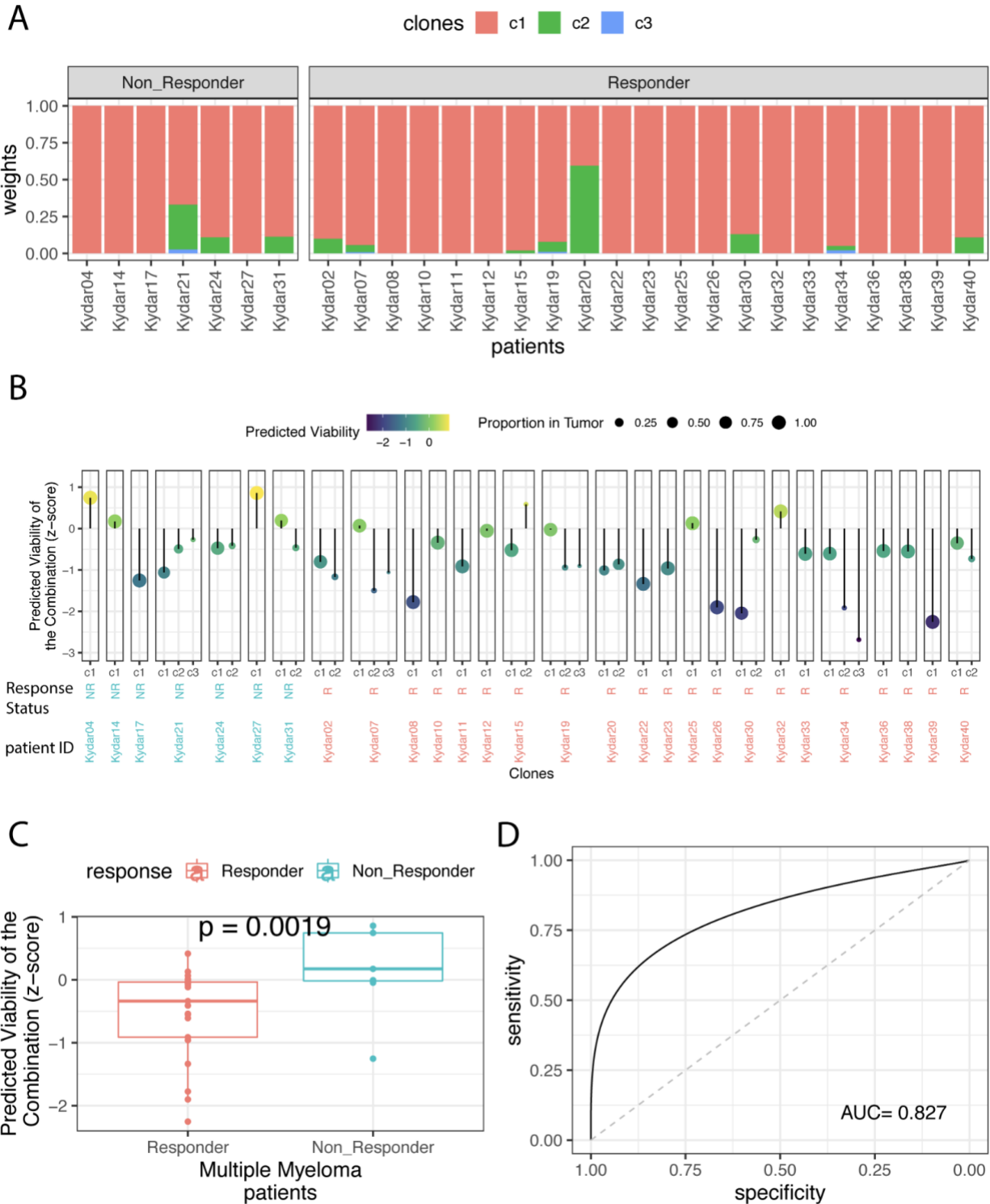


Figure 5: PERCEPTION stratifies responders vs non-responders of the combination DACA-KRD therapy regime in multiple myeloma patients. A) Distribution of abundance of malignant

sub-clones (y-axis) in each multiple myeloma patient (x-axis) in the trial identified using sc-expression, where the color code for the sub-clones is provided at the top. One-tailed Wilcoxon rank-sum p-value denoting the significance of the higher median predicted viability in resistant cell lines. B) Predicted viability of the combination (z-score) at a clonal level for each patient where response status is provided at the top-strip of each facet. The left to right order of patients is the same as in panel A. C) The predicted combination response in 28 multiple myeloma patients stratified by responders vs non-responders status. D) Receiver Operating curve based on predicted combination response shows the relationship between specificity (x-axis) and sensitivity (y-axis) is provided. The area under this curve, provided at the right bottom corner, denotes the stratification power to distinguish responders vs non-responders.

Charting the drug combinations landscape of FDA-approved drugs in multiple myeloma

After showing *PERCEPTION*'s ability to stratify myeloma patients to the DARA–KRD treatment, we next study its application for identifying combination treatments of FDA-approved cancer drugs for the multiple myeloma patients studied in the above trial that targets multiple clones in the tumor disjointly and thus, have a low likelihood for resistance emergence (Figure 6A). To this end, we first identified the set of FDA-approved cancer drugs that have predictive *PERCEPTION* models (N=44). We then searched for drug combinations that kill disjoint sets of clones in the tumor i.e. those that achieve maximal tumor coverage and killing (**Figure 6A**). We began with combinations of two drugs (946 possible pairs), ranking every pair by a score denoting the extent of their disjoint killing, termed its *Improvement Score* (IS). This score quantifies the fold increase in killing compared to expected (**Methods, Figure 6B**) predicted to be induced by the specific combination. Out of the 946 possible combinations scanned, 842 pairs show no improvement over the expected (Methods). The remaining combinations, with

Improvement score > 1 , termed *effective*, are shown in **Figure 6B**. The extent of the killing of different tumor clones by a few top-ranked effective combinations is shown in **Figure 6C** including our top-ranked combination hit of gefitinib & ponatinib, an EGFR inhibitor, and a canonical BCR-ABL inhibitor, respectively (IS = 2.57, Empirical P value = $1\text{E-}04$). Another top combination pair with the high improvement score is vinblastine and sunitinib (Improvement score = 1.55, Empirical P value = $1\text{E-}04$), a tubulin polymerization and multi-targeted receptor tyrosine kinase inhibitor, respectively.

Analogously, we next looked for all possible triplets of drug combinations (**Figure 6D**, **N=13,244**), where our top hits include the combination of etoposide + midostaurin + ponatinib (Improvement score = 2.09, Empirical P value = $1\text{E-}04$) and etoposide + niraparib + ponatinib (Improvement score = 1.95, Empirical P value = $1\text{E-}04$) (**Figure 6E**). This method can be utilized for optimal combination design targeted multiple clones in multiple myeloma patients where combinations with high improvement scores.

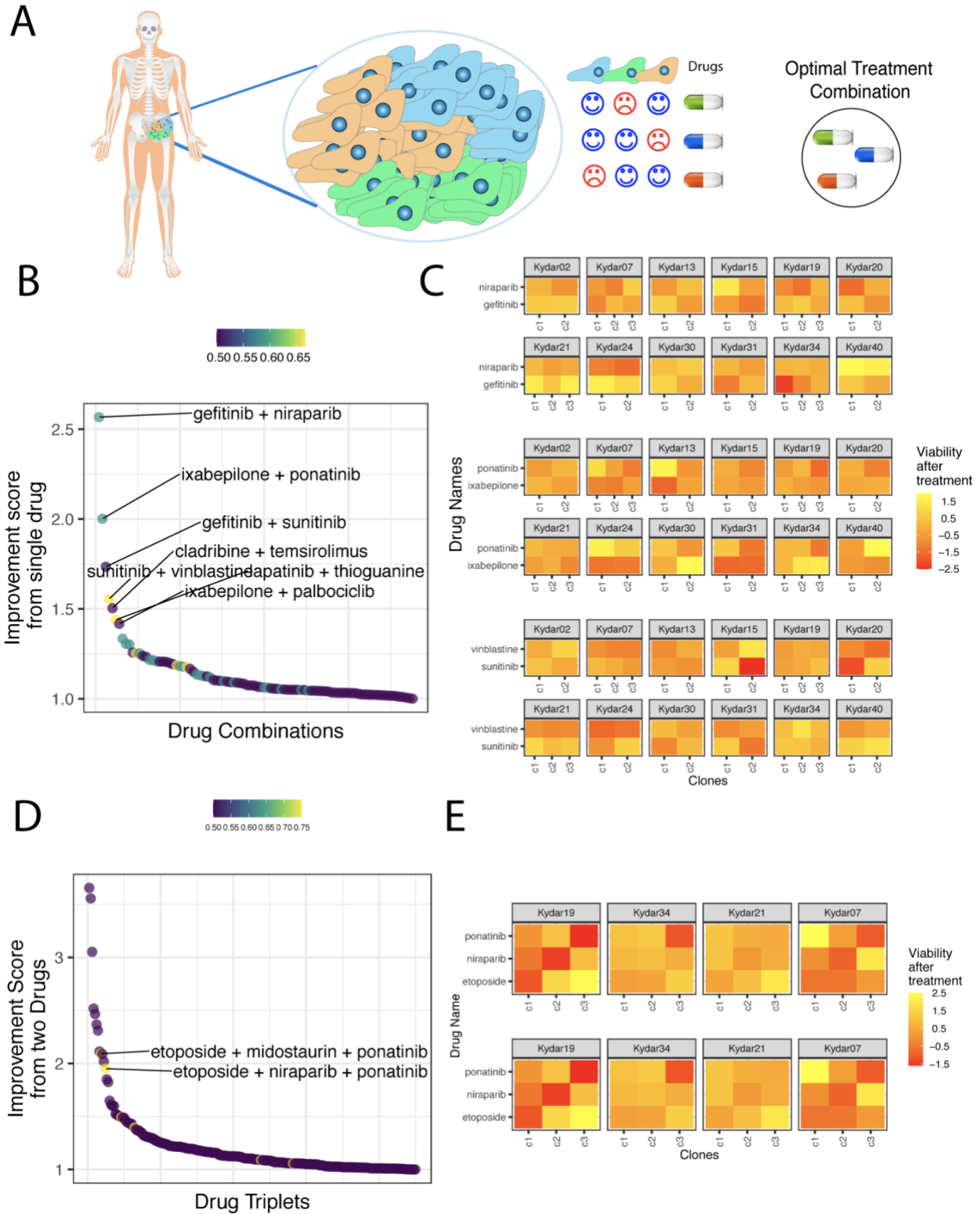


Figure 6: Identify optimal drug combinations for multiple myeloma patients. A) Overview of our method: Given a patient with multiple myeloma composed of three clones, our goal is to find

*a drug combination that can disjointly kill separate clones, and thus their combination treatment can kill all the clones present and could be an optimal combination. **B)** For each pair (x-axis), a median improvement score (y-axis) is computed i.e. fold decrease of predicted viability compared to the baseline. Here, an improvement score is provided for 94 pairs where this score is >1 . For each pair, the proportion of patients ($N=12$) where this score is greater than 1 is denoted by the intensity of the color, where the legend is provided at the top. Top pairs are labeled. **C)** Disjoint killing of clones is visualized using viability profile at a clone level for the top pairs from panel **C**. This is provided separately for each patient (a facet) where the intensity of the color denotes the viability after the treatment for each clone (x-axis) of a given drug (y-axis) where the legend is provided on the right. Panel **D)** and **E)** **are** analogous to panel **B)** and **C)**, respectively, but for drug triplets.*

Discussion

We present a framework to build drug response models to predict the drug response in cancer cells at single-cell resolution and demonstrate its application to predicting response to monotherapy and combination treatment at the level of cell lines, patient-derived-cell lines, and a clinical trial of multiple myeloma. The power of our analysis is limited by the availability of sc-expression patients' datasets with response labels. Consequently, we expect that our model would further refine and improve as such datasets would become more available. The current response signatures identified by *PERCEPTION* are pan-cancer and could be further refined by considering cancer type-specific cell lines, whenever a reasonable number of cell lines become available. We note that the quality of our response model would depend on the quality of the sc-expression profiles available e.g., depth, high drop-out rates, etc., and likely negatively impact our performance. To note, we chose not to impute our data after considering the recent progress

on the subject concluding that unexpected high-dropouts are limited to non-UMI-based sc-expression methods and likely due to biological variation (Svensson et al. 2020, Cao et al. 2021). The lack of toxicity and side effects screens in normal cells hindered us from learning these phenotypes. More availability of such screens in the future may enable our models to predict response to drugs in both tumor and normal cells and thus refine our combinations considering their toxicity as well. In summary, this study is the first to harness the high resolution of information from scRNA-seq technology to build drug response models that can be translated to clinical context to identify responders to combination therapy or design an optimal combination for an individual patient.

Methods

Data collection

We first collected the bulk-transcriptomics and drug response profiles generated in cancer cell lines curated in the DepMap (Tsherniak et al. 2017) consortium from Broad Institute (version 20Q1, <https://depmap.org/portal/download/>). The drug response is measured via area under the viability curve (AUC) across eight dosages and measures via a sequencing technique called PRISM (Corsello et al. 2020). In total, we mined 488 cancer cell lines with both bulk-transcriptomics and drug response profiles. We next mined sc-expression of 205 cancer cell lines (280 cells per cell line) generated in Kinker et al. 2020 from the Broad Single-cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP542/pan-cancer-cell-line-heterogeneity#study-download). The metadata, identification, and clustering information were also mined from the same portal.

The *PERCEPTION* pipeline

For each FDA-approved drug (N=133), we run the following steps to build a response model. **Step 1: Learn from Bulk** *A. Identifying gene bulk-expression features correlated with viability profile:* We first divided cell lines available to us into two sets, where the first set is used in step 1 and set 2 is used in step 2. We chose the cell lines where sc-expression is not available (N=318) and available (N=170) as sets A and B, respectively. Considering the first set of cell lines, we computed a Pearson correlation strength between the expression of each gene and for a query drug *d*. We considered this score as a measure of information in a gene expression profile and ranked each gene based on the correlated magnitude. *B. Build models:* Considering the top *X* gene expression features (where *X* is a hyperparameter optimized later) ranked based on the above-calculated correlation magnitude, we built a linear model regularized via the elastic net in five-fold cross-validation. **Step 2: Optimize using sc-expression.** *A. Hyperparameter Optimization using sc-expression:* We built the above model using a Bayesian-like grid of various *X* values (range 10-500), where the model with the best performance using sc-expression input of 169 cell lines (left one out for testing) is chosen. *B. Performance in Cross-Validation:* In the left-out cell line, which has not been used in either model building or hyperparameter optimization, we perform the error estimation. The error estimation is done by repeating the above steps for 170 times, wherein each instance a different cell line is left out. The final predicted values in this leave-one-out testing are compared to observed values via Pearson Correlation.

Quality control of in-house pharmacological screen by comparing them to existing screens

To test the quality of our *in-house* screen generated, we compared our screen to a previous high-quality screen from Broad and Sanger Institute, PRISM (Corsello et al. 2020). Specifically, we leveraged the fact that screens for these drugs are also performed in the same cell lines, at least their monotherapy. For each drug, we computed a correlation between our *in-house* screen and PRISM in matched cell lines. We reasoned that the drugs with correlated profiles in the two screens (Pearson $Rho > 0.3$) are consistent across the two screens suggesting that they are high quality. Independently, we note that the concordance score of drugs' response profile across screens is correlated with our predictive performance (Pearson $Rho > 0.39$; $P < 0.019$, **Extended Figure 2A**), suggesting that our model is capturing the robust signal across screens of these drugs.

Cross-platform comparison of PERCEPTION *performance*

The pharmacological drug screens performed by PRISM and GDSC studies are based on two independent platforms. The GDSC data was downloaded from the DepMap portal (Downloaded: April 15, 2020, <https://depmap.org/portal/download/>). To compare the performance of PERCEPTION across two independent screening platforms and test if the expression signature captured by our drug response models can be translated across the domains: 1. Of the 347 cell lines in common with drug response in GDSC and PRISM, there are 120 cell lines with sc-expression data. We selected at random 80 cancer cell lines with sc-expression data and pharmacological screens in GDSC and PRISM, 2. We considered all the drugs (N=191) which were screened in both PRISM and GDSC, from

which we selected a subset of drugs (N=28) with a concordant response between PRISM and GDSC (Pearson $\rho > 0.3$ and p-value < 0.05 ; at least 20 cell lines with responses per drug in both GDSC and PRISM) in the 267 cell lines in common between the two screens excluding the cell lines in the testing set. 3. For each of the selected drugs we ran the PERCEPTION pipeline, in Step 2 of the pipeline the parameters were optimized on sc-expression of 90 cell lines (excluding the 80 test cell lines) instead of the default 170 cell lines. 4. Finally, we applied the resulting response models to the testing dataset and compared the predicted AUC values to the experimental response from GDSC and PRISM. We used the Pearson correlation coefficient as the measure to compare the performance between the screens and predicted responses.

Generating models for PDC cell lines

The single-cell expression of the five HNSC patient-derived cell lines and their treatment response for eight drugs and combination therapy at two different dosages obtained from Suphavilai et al. 2020. For these drugs, PERCEPTION was unable to build drug response models using PRISM screens. Therefore, we incorporated two main changes to the PERCEPTION pipeline: 1. drug response from GDSC screens (response from $> \sim 800$ cell lines for these drugs) were used to build models, 2. the expression levels of all the genes in the cancer cell line datasets were not found in the PDC sc-expression datasets and the frequency of dropouts in the PDC dataset is higher— ~ 3500 genes have $> 75\%$ non-zero counts across all the cells in all five patients, as a result only the top 3000 genes which are common in both datasets and with fewer dropouts across all five patients are considered in the pipeline. For the drugs for which

PERCEPTION was able to build models, we applied the models on the PDC cell lines and obtained the predictions for each individual cell. The patient-level monotherapy response for a given drug is represented by the mean response of all the cells included in a patient's PDC sample. In the case of drug combinations, for a given cell, its combination response was the minimum among the two drug responses for the cell. The patient-level combination response was represented by the mean of the combined response of all the cells in a patient's PDC sample.

Data availability

The entire collection of the processed datasets used in this manuscript, including pre-clinical models of cancer cell lines and PDCs, can be accessed via a ZENODO repository which could be provided upon request or upon publication.

Code availability

We used open-source R version 4.0 to generate the figures. Wherever required, commercially available Adobe Illustrator 23.0.3 (2019) was used to create the figure grids. All of the scripts for analysis and figure production were built in-house and will be provided upon publication.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health, NCI. This work used the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). We acknowledge and thank the National Cancer Institute for providing financial and infrastructural support.

Conflict of interest

The authors declare that they have no conflict of interest.

Author contributions

References

1. Corsello, Steven M., et al. "Discovering the anticancer potential of non-oncology drugs by systematic viability profiling." *Nature cancer* 1.2 (2020): 235-248.
2. Kinker, Gabriela S., et al. "Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity." *Nature Genetics* 52.11 (2020): 1208-1218.
3. Ghandi, Mahmoud, et al. "Next-generation characterization of the cancer cell line encyclopedia." *Nature* 569.7757 (2019): 503-508.
4. Shalek, Alex K., and Mikael Benson. "Single-cell analyses to tailor treatments." *Science translational medicine* 9.408 (2017).
5. Adam, George, et al. "Machine learning approaches to drug response prediction: challenges and recent progress." *NPJ precision oncology* 4.1 (2020): 1-10.
6. Zhu, Sibio, et al. "Advances in single-cell RNA sequencing and its applications in cancer research." *Oncotarget* 8.32 (2017): 53763.
7. Kim, K.-T. et al. Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. *Genome Biol.* 17, 80 (2016).

8. Suphavitai, Chayaporn, et al. "Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures." *bioRxiv* (2020).
9. Cohen, Yael C., et al. "Identification of resistance pathways and therapeutic targets in relapsed multiple myeloma patients through single-cell sequencing." *Nature Medicine* 27.3 (2021): 491-503.
10. Ledergor, Guy, et al. "Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma." *Nature medicine* 24.12 (2018): 1867-1876.
11. Sade-Feldman, Moshe, et al. "Defining T cell states associated with response to checkpoint immunotherapy in melanoma." *Cell* 175.4 (2018): 998-1013.
12. Garnett, Mathew J., et al. "Systematic identification of genomic markers of drug sensitivity in cancer cells." *Nature* 483.7391 (2012): 570-575.
13. Tian, Luyi, et al. "Clonal multi-omics reveals Bcor as a negative regulator of emergency dendritic cell development." *Immunity* (2021).
14. Tsherniak, Aviad, et al. "Defining a cancer dependency map." *Cell* 170.3 (2017): 564-576.
15. Wensink, G. Emerens, et al. "Patient-derived organoids as a predictive biomarker for treatment response in cancer patients." *NPJ precision oncology* 5.1 (2021): 1-13.
16. Yao, Ye, et al. "Patient-derived organoids predict chemoradiation responses of locally advanced rectal cancer." *Cell stem cell* 26.1 (2020): 17-26.

17. de Witte, Chris Jenske, et al. "Patient-derived ovarian cancer organoids mimic clinical response and exhibit heterogeneous inter-and inpatient drug responses." *Cell reports* 31.11 (2020): 107762.
18. Cao, Yingying, et al. "Umi or not umi, that is the question for scRNA-seq zero-inflation." *Nature Biotechnology* 39.2 (2021): 158-159.
19. Svensson, Valentine. "Droplet scRNA-seq is not zero-inflated." *Nature Biotechnology* 38.2 (2020): 147-150.
20. Ling, Alexander, and R. Stephanie Huang. "Computationally predicting clinical drug combination efficacy with cancer cell line screens and independent drug action." *Nature communications* 11.1 (2020): 1-13.
21. Castro, L. Nicolas Gonzalez, Itay Tirosh, and Mario L. Suvà. "Decoding Cancer Biology One Cell at a Time." *Cancer Discovery* 11.4 (2021): 960-970.
22. Arya, Arvind K., et al. "Nutlin-3, the small-molecule inhibitor of MDM2, promotes senescence and radiosensitises laryngeal carcinoma cells harbouring wild-type p53." *British journal of cancer* 103.2 (2010): 186-195.

Conclusion

In this thesis, I provide an overview of the three computational approaches I developed during my Ph.D. to advance precision medicine for cancer prevention and treatment. The common axis among these three efforts is that we analyze large-scale cancer omics data from both pre-clinical models and patients datasets to generate the initial hypothesis and guide the study.

We leveraged large-scale genetic screens in cancer cell lines to identify the cancer risk associated with CRISPR-based therapies i.e. an undesired selection of cells with pre-existing p53 and KRAS mutations and thus calling for carefully monitoring patients undergoing CRISPR-Cas9-based editing for clinical therapeutics for pre-existing p53 and KRAS mutations.

We integrated publicly available and our own generated large-scale omics data from African American and European American cancer patients to delineate some of the molecular mechanisms that may underlie the observed differences in cancer incidences across cancer patients from these two ancestries. Focusing mainly on lung cancer, we found that lung tumors from African American (AA) patients exhibit higher genomic instability, homologous recombination deficiency, and aggressive molecular features such as chromothripsis. These molecular differences extend to other cancer types. We also showed that these somatic differences observed may have genetic origins by comparing germline variants from patients of these two ancestries. This provides a therapeutic unique strategy to treat tumors from an ancestry (AAs) with high HRD using PARP and checkpoint inhibitors.

Finally, I presented a computational framework to use single-cell RNA-seq from patients' tumors to find combination treatments that can target multiple clones in the tumor disjointly. Using this framework, we predict the response to monotherapy and combination treatment in cell lines, patient-derived-cell lines, and in a clinical trial of multiple myeloma patients and chart the landscape of optimal combination treatments of the existing FDA-approved drugs in multiple myeloma.

In summary, we demonstrate the power and wide range of applications of multi-omics analysis to identify cancer risks associated with genetic editing and strategies to overcome it, to delineate molecular mechanisms contributing to the cancer disparity in AA patients, and a treatment strategy based on that, and finally, we built a framework based on single-cell transcriptomics of tumors to stratify responders and guide combination treatment.